



Variability in teacher oral English language assessment decision-making

Author:

Phung, De

Publication Date:

2018

DOI:

<https://doi.org/10.26190/unsworks/20467>

License:

<https://creativecommons.org/licenses/by-nc-nd/3.0/au/>

Link to license to see what you are allowed to do with this resource.

Downloaded from <http://hdl.handle.net/1959.4/60023> in <https://unsworks.unsw.edu.au> on 2024-04-19

VARIABILITY IN TEACHER ORAL ENGLISH ASSESSMENT

**Variability in teacher oral English language assessment
decision-making**

De Van Phung

A thesis in fulfilment of the requirements for the degree of Doctor of
Philosophy



School of Education

Faculty of Arts and Social Sciences

June 2018

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family Name: Phung

First Name: De

Other Name/s: Van

Abbreviation for degree as given in the University calendar: PhD

School: School of Education

Faculty: Faculty of Arts and Social Sciences

Title: Variability in teacher oral English assessment decision-making
Abstract

Assessment decision-making is an integral part of teacher practice. Issues related to its trustworthiness have always been a major area of concern, particularly variability and consistency of judgment amongst teachers. There has been extensive research on factors affecting variability, but little is understood about the cognitive processes that impact the trustworthiness of assessment. Even in an educational system like Australia, where teacher-based assessment in the mainstream schooling system is widespread, it has only been relatively recently that there have been initiatives to enhance the trustworthiness of teacher assessment of English as a second or additional language or dialect (EAL/D), but how teachers make their decisions in assessing student oral language development has not been documented. In this study, I explored this issue using the oral assessment tasks and protocols developed as part of the Victorian project, Tools to Enhance Assessment Literacy for Teachers of English as an Additional Language (TEAL). I adapted the materials and applied them in the context of EAL/D learning and teaching in New South Wales, aiming to (1) examine to what extent EAL/D teachers' oral assessments were consistent, (2) explore factors influencing their assessments, and (3) identify characteristics of teacher decision-making. Employing a mixed-method research approach, this study involved twelve experienced NSW primary and secondary EAL/D teachers who participated in a survey, an assessment activity and a think-aloud protocol followed by individual interviews. The findings revealed that teachers were different from each other in the ways in which they came to their judgment decisions and in their perception of student development and that the differences were affected by factors related to teacher and student demographics and the characteristics of the assessment tasks. One result of this study was the development of a new framework to understand teacher decision-making processes with three different styles namely: (1) self-regulated assessment, (2) conflicted assessment, and (3) automated assessment. These decision-making styles provide a new lens for explaining variability in teachers' judgement of student oral language development. Implications of the framework for assessment theory and practice, teacher development, policy articulation and future research are also discussed.

Declaration relating to disposition of project thesis/dissertation

I hereby grant permission to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350-word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctorate theses only).

Witness

Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY

Date of completion of requirements for Award

Acknowledgements

First, I would like to express my gratitude to my main supervisor Professor Chris Davison for her timely and continuous support, her patience, encouragement, advice and guidance during my four-year PhD candidature. Her guidance helped me throughout my research and the writing of my thesis. I would also like to give thanks to Chris for providing me with opportunities for professional development by sponsoring conference and workshop attendance. I could not have imagined having a better adviser and mentor for my PhD.

I would also like to thank Dr Michael Michell, who gave me valuable support and advice on the interpretation of analysed results and the writing of my thesis. In addition, my special thanks go to Dr Dennis Alonzo, who gave valuable comments and suggestions on my thesis writing. I would also like to thank him for his advice on securing and balancing my student life. I would like to thank Associate Professor Jihyun Lee, my former co-supervisor, for her valuable guidance and enthusiastic support during the first two years of my PhD candidature.

My thanks also go to the Department of Foreign Affairs and Trade of Australia for their full financial support during my study. I would like to thank Mr Matthew Byron, Australian Award Contact Officer, who gave me timely and helpful support and demonstrated patience during my candidature. My thanks also go to Tra Vinh University for granting me permission to take a long study leave.

My gratitude also goes to all the teachers who participated in this study. Their invaluable contribution has resulted in significant findings. Thank you to all my PhD student fellows who gave me practical and valuable feedback and support throughout my research.

Finally, I would like to express my thanks to my wife, who was always with me, had my back and took over all our family's responsibility during my candidature. I also dedicate this thesis to my family in Vietnam, especially my amazing parents, who give me constant moral support and encouragement throughout my life's journey.

Contents

Originality Statement.....	iv
Copyright Statement.....	v
Authenticity Statement.....	vi
Acknowledgements.....	vii
Contents.....	viii
List of Figures.....	xi
List of Tables.....	xii
List of Appendices.....	xiii
List of Abbreviations.....	xiv
Chapter 1. Introduction.....	1
1.1. Introduction.....	1
1.2. Teacher Language Assessment and Related Issues.....	2
1.2.1. Variability: a threat to the reliability of language assessment.....	5
1.3. My Motivation.....	10
1.4. Scope and Aims of the Study.....	12
1.5. Significance of the Study.....	13
1.6. Structure of this Thesis.....	14
Chapter 2. Literature Review.....	16
2.1. Introduction.....	16
2.2. Teacher-Based Assessment: Why, What and How.....	17
2.3. A Shift in Assessment Theory.....	19
2.3.1. Anderson's framework for alternative assessment.....	19
2.3.2. Brookhart's classroometric perspective of assessment.....	22
2.4. Variability in Teacher-Mediated Language Assessment.....	28
2.5. Previous Studies on Teacher Assessment Variability.....	32
2.5.1. Variability related to teachers' experience.....	32
2.5.2. Variability related to teachers' and students' gender.....	35
2.5.3. Variability related to teachers' main exposure to a language group.....	37
2.5.4. Variability related to teachers' educational backgrounds.....	43
2.5.5. Variability related to students' language background.....	44
2.5.6. Variability related to tasks.....	47
2.5.7. Variability related to scoring rubrics.....	48
2.6. Teacher Assessment Decision-Making.....	50
2.6.1. A framework for teachers' classroom assessment decision-making.....	52
2.6.2. Gestalt in teacher assessment decision-making.....	56
2.6.3. Flexibility and moderation in teacher decision-making.....	61
2.7. A Need for a Conceptual Framework for Teacher Language Assessment Decision-Making.....	68
2.8. Summary.....	71
Chapter 3. Methodology.....	72
3.1. Introduction.....	72
3.2. Research Approach.....	74
3.2.1. Pragmatism: a philosophical partner to mixed methods research.....	74
3.2.2. Mixed methods approach.....	75
3.2.3. Justification for mixed methods research.....	80
3.3. Context of the Study.....	82

3.3.1. Development of the Teacher-Based Oral Assessment System	83
3.3.2. EAL/D learning and teaching in NSW	84
3.4. Research Design	86
3.5. Stage 1	87
3.5.1. Participants	87
3.5.2. Questionnaire	90
3.5.3. Teacher-based assessment activity	91
3.5.4. Data collection procedures	95
3.5.5. Retrospective think-aloud activity	97
3.6. Stage 2	101
3.6.1. Participants	101
3.6.2. Retrospective think-aloud	101
3.7. Pilot Study	103
3.8. Data Analysis	104
3.8.1. Analysis of assessment data	104
3.8.2. Analysis of questionnaire data	112
3.8.3. Analysis of document data	112
3.8.4. Analysis of interviews and group discussion data	113
3.9. Ethical Considerations and Research Validity	113
Chapter 4. Variation in Teacher Scorings	115
4.1. Introduction	115
4.2. Variations for Individual Students	116
4.3. Variations of Individual Teachers Across Students	119
4.4. Variations for Individual Categories	124
4.5. Discussion	129
4.6. Conclusion	139
Chapter 5. Interactions in Teacher Assessment Decisions	141
5.1. Introduction	141
5.2. Interactions in Teacher Assessments	142
5.2.1. Quantitative interactions with background factors	143
5.2.2. Qualitative interactions with background factors	157
5.3. Interactions with Assessment Factors	169
5.3.1. Student-related factors	169
5.3.2. Assessment task-related factors	171
5.3.3. Assessment criteria-related factors	174
5.3.4. Scoring procedure –related factors	177
5.4. Conclusion	183
Chapter 6. Teacher Decision-Making Processes	185
6.1. Introduction	185
6.2. Self-Regulated Assessment Style	190
6.3. Conflicted Assessment Style	198
6.4. Automated Assessment Style	202
6.5. Comparisons of Teacher Assessment according to Assessment Styles	210
6.5.1. Demographic differences among assessment styles	210
6.5.2. Distribution of interactions with background and assessment-related factors	211
6.5.3. Disagreement in perceptions among teachers	213
6.5.4. Variability and consistency differences among groups	221
6.6. Conclusion	227

Chapter 7. Discussion.....	229
7.1. Introduction	229
7.2. A New Approach to Understanding Teacher Assessment Decision-Making.....	230
7.2.1. Using variability as a resource in teachers' assessment decisions.....	232
7.2.2. The role of flexibility and scoring rubrics in improving consistency in teacher assessment decisions	235
7.3. Moderation to Improve Consistency	242
7.4. Revisiting Trustworthiness	243
Chapter 8. Conclusions and Implications.....	245
8.1. Introduction	245
8.2. Summary of Key Findings	246
8.3. Implications.....	247
8.3.1. Implications for theory and practice.....	248
8.3.2. Implications for pre-service teacher training and professional learning	251
8.3.3. Implications for educational policy	253
8.4. Limitations and Suggestions for Further Research	254
8.5. Final Thoughts.....	256
References	258
Appendices.....	289

List of Figures

Figure 2.1. Teacher Classroom Assessment Decision-Making.....	55
Figure 2.2. Gestalt Principle of Perception Organisation	58
Figure 2.3. A subjective contour.	59
Figure 2.3. Framework for Teacher Reflection.....	63
Figure 3.1. Subtypes of Mixed Methods Research.	77
Figure 3.1. Research Design.	86
Figure 4.1. Variability across Students.	121
Figure 4.2. Consistency Stability across Students.	123

List of Tables

Table 3.1 <i>Teachers' Demographic Information</i>	88
Table 3.2 <i>Non-Scaled Deviations on Communication for SI</i>	107
Table 3.3 <i>Actual Absolute Deviations Across Students</i>	108
Table 3.4 <i>Scaled Absolute Deviations</i>	110
Table 3.5 <i>Sampled Modification of Absolute Deviations</i>	111
Table 4.1 <i>Variability for Individual Students</i>	117
Table 4.2 <i>Consistency for Students</i>	119
Table 4.3 <i>Variability for Individual Categories</i>	125
Table 4.4 <i>Consistency for Individual Categories</i>	129
Table 5.1 <i>Correlations Between Background Factors and Teacher Assessments</i> ...	146
Table 5.2 <i>Contribution of Background Factors to Variances of Teacher Assessments</i>	148
Table 5.4 <i>Correlations Between Background Factors and Assessment Categories: Variability</i>	153
Table 5.5 <i>Correlations Between Background Factors and Assessment Categories: Consistency</i>	156
Table 6.1 <i>Differences among Decision-making Styles: Interactions with Influential Factors</i>	212
Table 6.2 <i>Differences among Decision-making Styles: Variability and Consistency</i>	223

List of Appendices

Appendix A Participant Information Statement and Consent Form	289
Appendix B Questionnaire.....	294
Appendix C Performance Task Descriptions Adapted from TEAL.....	296
Appendix D Assessment Criteria	301
Appendix E Guidelines for Think-aloud Protocol and Interview Questions.....	309
Appendix F Participant Information Statement and Consent Form.....	310

List of Abbreviations

ACARA	Australian Curriculum Assessment and Reporting Authority
ATESOL	Association of Teachers of English to Speakers of Other Languages
CC	Correlation Coefficients
EAL	English as an Additional Language
EAL/D	English as an Additional Language or Dialect
EFL	English as a Foreign Language
ELL	English Language Learners
ESL	English as a Second Language
ETS	Educational Testing Service
HSC	Higher School Certificate
IELTS	International English Language Testing Service
MELAB	Michigan English Language Assessment Battery
NAPLAN	National Assessment Program—Literacy and Numeracy
NCME	National Council on Measurement in Education
NSW	New South Wales
RPD	Recognition-primed Decision
SERAP	State Education Research Approvals Process
TEAL	Tools to Enhance Assessment Literacy
TOEFL	Test of English as a Foreign Language
UK	United Kingdom
US	United States

Chapter 1. Introduction

1.1. Introduction

Effective teaching is known to be guided by teacher decision-making (Wilén, Bosse, Hutchison, & Kindsvatter, 2004). This is because teacher assessment decisions about their students' proficiency allow the teacher to reflect and evaluate the effectiveness of their teaching practice and decide on what and how possible teaching content and methodology actions will be taken. Teacher decisions inform students, as well as teachers, of students' strengths and weaknesses that they need to work on. Therefore, it has been suggested by many that effective learning assessment should be integrated with instruction (Shepard, 2000, 2001; Stiggins, 2002; Stiggins, Arter, Chappuis, & Chappuis, 2004). This suggestion adds weight to the importance attached to teacher assessment decision-making, an area that has been challenged by the concept of variability.

This chapter focuses on the importance that is attached to teacher assessment decision-making and the significance of conducting this research. Issues in teacher language assessment will first be presented, detailing characteristics of teacher language assessment decisions, the role of teachers and variability in teacher assessment decision-making. The scope, aims and significance of this research will also be introduced, followed by the motivation of the researcher in conducting this study, and the structure of the study.

1.2. Teacher Language Assessment and Related Issues

Although research into high-stakes English language testing and assessment in English as a first language, second language (ESL), additional language or dialect (EAL/D) and a foreign language (EFL) is well-established, research into teacher language assessment has been, for several reasons, regarded as less important. The first reason for this is that, in comparison with assessments for summative purposes that are commonly conducted by an external agency and, thus, high-stakes, teacher-based language assessment is usually confined to a range of informal assessment tasks and activities that are carried out in the classroom as an integral component of learning and teaching and are considered relatively low-stakes. Teachers use these informal assessment activities to identify students' strengths and weaknesses, and then use this information to facilitate learning and support instruction.

Another reason for the relative lack of research into teacher-based assessment is the dominance of large-scale standardised language tests, including the Test of English as a Foreign Language (TOEFL), the International English Language Testing Service (IELTS), the Test of English for International Communication and standardised tests of German, Japanese, French, Chinese and so forth, with priority given to research which has investigated almost all aspects of these tests. However, as Wiliam (2001, p. 167) notes: 'Why rely on an out-of-focus snapshot taken by a total stranger?'. Teachers better understand their students' capability than others far removed from the classroom. The benefit of using teacher assessment for facilitating learning has also long attracted considerable attention in many English speaking countries (Carless, 2010).

For these reasons, although not well-researched, teacher assessment practice is now supported by policy in many countries around the world (Davison, 2004). One example is Australia, in which school-based assessment is compulsory and is including

as an integral component of all school certification. Teacher judgements are key in school-based assessment (Cumming & Maxwell, 2004). Similarly, in a national assessment program in Scotland, Assessment is for Learning, teachers' judgements are an integral part of understanding and sharing standards in summative assessment nationally. In Hong Kong, according to the policy of the Hong Kong Examinations and Assessment Authority, teacher-based assessment is a part of the existing formal territory-wide examination system across most senior secondary subjects (Hamp-Lyons, 2009). In Mainland China, policy attention is increasingly being drawn to teacher assessment and it was recently included in the tertiary curriculum framework of EFL (Xu & Liu, 2009). Further, in some African developing countries such as South Africa (Pryor & Lubisi, 2002) teacher assessment is increasingly being employed as national education policy. In the USA, due to the lack of national testing programs, the role of teacher assessment is actively recognised and promoted (Popham, 2014). In the same vein, the Ministry of Education of Vietnam recently issued regulations on assessment in primary and secondary education. Specifically, English instruction is deployed from third-grade and assessment is implemented on a regular basis for learning purposes by classroom teachers.

Although many may consider it a breakthrough that teacher assessment is supported worldwide, especially in English language education, there remains reluctance and suspicion among educators, researchers and even teachers themselves. Essentially, teacher assessment involves teachers controlling assessment processes from the beginning to the end. For the purpose of monitoring and evaluating students' progress, language teachers need to be knowledgeable and skilled in designing and using their own assessment tools (Davison & Leung, 2009). However, as highlighted by Davison and Leung, teacher language assessment features 'much variability, a lack of

systematic principles and procedures, and a dearth of information as to the effect of teacher-based assessments on learning and teaching' (p. 394). This is particularly the case in teacher-based oral language assessment. Teacher-mediated oral language assessment

Learning to speak in English (a productive skill) is difficult and learning to communicate orally in the language is even more difficult, thus, assessing someone's ability to use that language is a complicated process.. Hamp-Lyons (2007) argues that to achieve high-quality oral assessment, teacher assessments are required to be fair and valid. To achieve these two assessment goals of fairness and validity, it is necessary to provide teachers with adequate opportunities for professional development and the opportunity to work collaboratively within and across schools to enhance teacher assessment literacy and competence. Hamp-Lyons proposes that quality assurance and validation in teacher assessment should be conducted on an ongoing basis.

In educational and assessment research validity and reliability are two vital constructs that teachers must always take into consideration when making quality assessments of students' work (McNamara, 1996, 2000; Popham, 2011). Validity refers to ensuring that what is supposed to be assessed is assessed by using an appropriate assessment instrument (Popham, 2011).

Teacher-based assessment is "non-standardized local assessment carried out by teachers in the classroom" (Leung, 2005, p. 871). In classroom contexts, teachers are the main agent of, or central to, the entire assessment process (McNamara, 1996). Their assessment tasks should reflect the objective of a course or a unit of teaching, and need to be as authentic as possible to suit students' changing needs. Teachers also need to interpret their students' performances and align their interpretations with assessment standards. It should be noted that assessment standards, also referred to as a rubric or

criteria, for one oral task may not be appropriate for another. It is the teacher's job to select the rubrics or criteria to be used for assessing performance on a task type, then interpret and use the criteria to make decisions about their students' work.

Much effort has been made to theoretically assist teachers in their assessment decision-making in their classrooms (Borko & Shavelson, 1990; Colton & Sparks-Langer, 1993; McMillan, 2003; Westerman, 1991). Through conceptualising teacher beliefs and views and their assessment practices, assessment principles and guidelines have been developed to provide teachers with sufficient knowledge and a deeper understanding of what assessment decision-making looks like, what factors influence the process and what should be done to increase the quality of assessment decision-making practices. However, such efforts to ensure consistency in teacher assessment are challenged. Given that subjectivity is inherent to teachers (McNamara, 1996, 2000), their involvement in the process of assessing may result in variability (also known as inconsistency or assessment bias) in their assessment decisions that could threaten assessment trustworthiness.

1.2.1. Variability: a threat to the reliability of language assessment

As defined by the term itself, teacher language assessment is conducted by classroom teachers with their own students. Teachers are involved in the entire process of assessing their students, from planning and designing assessment instruments to scoring and then reporting the results, even when they are not assessment literate, or when quality assessment resources are not available to them. As noted by Gu (2014), 'Without basic training in assessment literacy, the curriculum mandate of formative assessment will definitely remain on paper only, no matter how many exemplars are provided' (p. 301)..

Lack of assessment literacy and standardised assessment tools may place pressure on teachers, perhaps driving them to adopt more idiosyncratic approaches they think are suitable to judge students' language development. Hence, teacher-based assessment is considered a source of variability (Davison & Leung, 2009), or assessment bias (Popham, 2014). The issue of variability is further complicated by the common view that teachers cannot help but be subjective when they judge their own students. Variability in oral language assessment is assumed to occur when a teacher's assessment of a student's language level is influenced by several different factors (Cooksey, Freebody, & Wyatt-Smith, 2007; Popham, 2004; Wyatt-Smith, 1999; Wyatt-Smith & Castleton, 2004). Davison (2004) argues that these influential factors arise from variation in teacher assessment beliefs, attitudes and practices. From another perspective, while some variability might be due to teacher background, much is related to variables intrinsic to the assessments themselves such as criteria, setting and tasks.

1.2.1.1. Teacher differences in classroom assessment

Although teachers as classroom assessors have long been known as a source of variability, teachers play vital roles in any sort of assessment for whatever purposes in their classroom. Teachers' individual differences may contribute to their variability when assessing their students. The fact that teachers differ from one another is obvious. Teachers' differences, individual and contextual, may affect their assessment practice in classrooms. While individual differences are associated with background factors such as age, gender, teaching experience or education and language background, contextual differences include the level of teaching and the main language group with which they work.

As a result of individual and contextual differences, for example, when assessing student oral language performances, teachers focus on both linguistic and non-linguistic aspects of their performances (Butler, 2009). Such differences also cause variability among teachers in scoring individual performances and create different understanding of the criteria. Individual differences and contextual differences together means that the assessment decision-making process will vary between teachers.

Teacher engagement with assessments materials and tools also vary. While some teachers report that their priority is placed on utilising objective assessment tools, others report that they pay more attention to using prompts (McMillan & Nash, 2000) to assess student abilities. Some teachers focus more on criteria, others rely on their ‘gut instinct’ (Davison, 2004). Hence, variability has emerged as concern in teacher-based assessment, exacerbated by increasing diversity among students coming from different cultural backgrounds. This diversity creates challenges for researchers, educator, teachers and other stakeholders to face in preparing for those students to enter the mainstream schooling system.

1.2.1.2. Resolutions for variability in language assessment

In many countries, great attempts have been made at national and state levels to alleviate the problem of variability in language assessment. A typical instance is the case of Australia. English is the instructional medium in all Australian schools, hence it is used as a medium to assess learning and achievement. All students including students with English as an additional language or dialect (EAL/D) take English as a core subject (AusVELS, 2013). EAL/D students come from diverse cultural and educational backgrounds and speak languages other than English. Therefore, they need additional support to develop their English oral communication skills as well literacy. While some

may view these students as a great source of diversity for enhancing classroom dynamics (Drucker, 2003), others express concern that a considerable proportion of these students will not be able to catch up with their English speaking peers in mainstream schooling system in less than five years (Cummins, 1996). With the commitment to providing all students with equal access to the Australian curriculum, the Australian Curriculum Assessment and Reporting Authority (ACARA) provides an EAL/D resource to support teachers with their EAL/D teaching from foundation to Year 10. EAL/D students then take the same English program as their native peers in upper secondary schools in years 11 and 12.

As identified in the EAL/D Overview and Advice Brochure (ACARA, 2014), for assessment, EAL/D teachers are required to be sufficiently knowledgeable and skilled to assess EAL/D students' language development and identify their learning needs and help them access the curriculum across all key learning areas. It is important to implement diagnostic assessment to support such teaching, and formative and summative assessment are also required in order to develop students' language skills, including listening, speaking, reading and writing. However, this source of support for EAL/D assessment is a general guideline and, therefore, each state or territory in Australia develops its own, more detailed and appropriate assessment systems for EAL/D education.

As an example, the Victorian Curriculum and Assessment Authority identifies the pathways to achieve the goals of EAL/D students as different from those students who speak English as their first language. Taking this into account, EAL/D goals and standards must be mapped onto the English standards. Accordingly, there are three EAL/D learning stages: A, B and S that are mapped against 11 levels of AusVELS achievement standards. In addition, AusVELS (2013) presents the EAL/D Companion

to AusVELS which provides a framework to assess EAL/D students and the English progress of EAL/D students is reported against EAL/D standards (e.g., A, B and S stages) rather than English standards. Their achievement is reported against the English standards only when the assessment of EAL/D students on the English standards is in an acceptable year level.

However, despite these guidelines for reporting on EAL/D development, Davison and Michell (2014) in an analysis of the assessment needs of teachers in Victoria and NSW, two of the most diverse and populous states in south-eastern Australia, found a complete lack of EAL/D assessment resources, and a tendency among teachers to develop and use their own assessment tasks or adapt other teacher-developed assessment materials. Those teachers who use home-grown assessment materials are often concerned about the quality and the reliability of their assessments.

In response to this urgent and very practical demand for improvements in the quality of assessment advice and resources, a project to develop more standardised ESL assessment tools and advice, undertaken by a research team at the University of New South Wales (UNSW), was commissioned by the Victorian Department of Education and Early Childhood Development, the Catholic Education Commission of Victoria and Independent Schools in Victoria, and officially launched in 2015. The goal of this project, called Tools to Enhance Assessment Literacy for teachers of English as an Additional Language (TEAL), see <http://teal.global2.vic.edu.au/>, project was to produce an online assessment 'toolkit' developed by EAL/D teachers for EAL/D teachers in their own classrooms to improve their assessment literacy as well as their confidence and trustworthiness. The TEAL project consisted of three main components: 1) development of the web-based assessment resource centre, 2) development of the prototype teacher-

based writing and oral assessment system and 3) development of the computer adaptive test of reading and vocabulary.

This study aims to draw on the TEAL assessment resources (specifically the system of teacher-based oral assessment publicly available on the website and described more fully in Chapter 3), to provide more knowledge and understanding of variability and the processes teachers use to make assessment decisions.

While most of the studies on variability in teacher assessment explore it through a psychometric lens, this study aims to investigate variability in teacher decision-making through a 'classroometric' perspective (Brookhart, 2003). Assuming that variability is an inherent characteristic to human assessors (Davison & Leung, 2009) and teacher assessment or classroom assessment is construct- and -context-dependant , variability in teacher assessment decisions is unavoidable and, thus, should be exploited rather than eliminated to improve assessment practice and support teaching and learning. By adapting the assessment resources developed for the larger TEAL project, this study seeks to observe how experienced teachers with minimal training use unfamiliar resources to assess the oral language skills of students they do not know. Given the limitations of more traditional psychometric approaches to examining teacher variability, discussed in more detail in Chapter 2, this study approached teacher assessment decision-making through a different perspective drawing on Gestalt theory (Wertheimer, 1912, as cited in Wertheimer, 2012).

1.3. My Motivation

My prime motivation for this research comes from my professional background in my home in Vietnam where, like Australia and other countries, variability in teacher-based English language assessment exists, but is little understood. As an instructor who

teaches English to students from different disciplines and to English major students at a newly established university in the south of Vietnam, I find that the nature of my teaching is somewhat like teaching English to newly arrived students in the Australian schooling system.

As required by the curriculum, I have made a wide range of day-to-day assessments of my students' oral communication skills without having received any formal training in assessment approaches. I have made assessments of this kind for several years and have trained myself by learning from my own year-by-year accumulated experience as well as from my senior colleagues, through personal interaction and by adapting existing assessment tasks from commercial sources or other teacher-made assessment materials and instruments. General English courses are compulsory and taught in the first seven semesters of a four-year program and five semesters of a three-year program. As for English majors, four main language skills are taught separately. Evaluation of a course consists of 50 per cent of ongoing assessment including student attendance and mini assessment tasks, with the remaining 50 per cent derived through a final test. Both the ongoing assessment tasks and the final test are designed and assessed by the teacher of that course. The teacher then uses all the assessment information to make their final decisions on student performance of the course (i.e., passing or failing).

There are few opportunities for teacher's assessment practice to be shared or reviewed; therefore, it is almost unknown whether their assessment practice is accurate or consistent. Classrooms can be viewed as black boxes in which the actions teachers from different backgrounds take to assess students is a well-kept secret. Further, it is not compulsory for teachers to share or discuss their assessments with each other. While there are still some teachers who are willing to share and raise discussions with others,

this sometimes happens to satisfy some personal demands. The only opportunity for teachers to cooperate in assessment is when two or three teachers are assigned to score graduation papers or presentations. This is the only occasion when teachers' assessments are compared and reviewed. Every time this happens, differences among teachers are always observed and dealt with through a process of moderation. I am particularly interested in exploring teachers who, like myself, assess their students' oral English communication skills when a standardised assessment instrument is not available and when they are not quite confident with their assessment competence and literacy. New knowledge and deeper understandings of how teachers perceive student work, interpret standards and locate student language development on a proficiency continuum will help me to improve my assessment and teaching practice and support my students' learning.

1.4. Scope and Aims of the Study

As stated earlier, teacher assessment in language education needs to be fair and valid. Variability in teacher assessment decision-making may occur regardless of the contexts in which assessment is carried out, because the nature of teacher assessment decision-making is subjective. Therefore, it is necessary to not only conduct research to explore what teachers do to assess students' language work, but also to identify the factors which influence their assessments. Teacher assessment of students' work may be driven by many factors; however, this study is designed to explore factors in relation to the teachers themselves, their students and the assessment tasks. This study first seeks to determine the consistency of teachers' assessments. At the same time, this study also aims to explore the factors that contribute to variability in teachers' assessments, as well

as the characteristics of teacher decision-making. The study will address the following questions.

1. To what extent are teachers' assessments of students' oral English communication skills consistent with one another?
2. What are the factors that influence teachers' assessments?
 - a. What factors related to teachers' background influence their assessments?
 - b. What factors related to the assessment tasks affect teacher's assessments?
3. What are the characteristics of teacher assessment decision-making?

1.5. Significance of the Study

Conducting this study into teacher oral English language assessment is of significance in the following ways. First, teacher-based English language assessment has not been well documented in the research literature and variability in teacher oral language assessment decision-making have yet been investigated in ways which are congruent with a new perspective on teacher-based assessment. Therefore, it is important to explore how consistently teachers assess their students' oral communication skills, and to discover what factors influence their assessments in diverse ESL/EAL/D classrooms, not only in Australia but across the world. The aim is to provide an alternative approach to unearthing the process of teacher assessment decision-making. Findings from this study can be conceptualised into guidelines that will help teachers to develop their assessment knowledge and skills and be operationalised in classrooms to improve teacher assessment practices.

In addition, conducting this study is also globally significant, as given the rise of teacher-based assessment, even in traditional examination cultures, the issue of

variability in ESL or EAL/D assessment is now a worldwide concern, with psychometric models of assessment quality often the only source of evaluation. Therefore, assessment conceptions or guidelines that develop from the findings of this study may be useful and applicable across a range of different contexts in which English is taught as a second or additional language. Those conceptions and guidelines can be adapted to improve the quality of the assessment system in Vietnam in general and in my institution, Tra Vinh University, where English is taught as a foreign language.

Finally, since the TEAL project was the starting point for this study, the findings will help evaluate its assessment guidelines and resources and contribute to measuring the effectiveness of the project. The findings of this study will be also useful as the assessment tools developed by TEAL in this study were used by experienced teachers from other assessment districts who had minimal training in using those tools. Findings from this study can show whether what has been created by TEAL can be used or reproduced in other states or territories. To conclude, for those reasons stated above, this research study is significant on multiple levels.

1.6. Structure of this Thesis

In this chapter, I have introduced the importance of teacher assessment decision-making and identified issues of concern in this area. I have also identified various aspects of the research problem which have been operationalised into research questions. Then, the contributions this study can make were presented, followed by my rationale and personal motivation to conduct this study. In Chapter 2, I review the relevant literature regarding variability in teacher oral language assessment and the theoretical literature shaping teacher decision-making practice. Also, in this chapter, I explain the need to have a more relevant conceptual framework to support teacher

assessment decision-making. In Chapter 3, I discuss and justify the mixed method approach adopted in this study, followed by the design of this study explaining how data are collected and analysed. In Chapter 4, I examine the differences among teachers' assessments in terms of variability and consistency along three different dimensions, followed by the findings regarding factors influencing assessment practice, which are presented in Chapter 5. In Chapter 6, I present one of the major findings of this study, which is the conceptual framework explaining how teacher assessment decisions are made. Chapter 7 discusses and explains the entire process of teacher decision-making and its contribution to the literature. I also discuss the role of assessment criteria, flexibility and moderation in improving consistency in teacher assessment decision-making, followed by the role of trustworthiness in teacher assessment decision-making. I conclude this thesis by summarising the major findings, followed by implications for theory and practice, teacher training and educational policy practices. I also present the limitations of this study by embedding them in suggestions for further research.

Chapter 2. Literature Review

2.1. Introduction

Chapter 1 presented the background and rationale for this study. Chapter 2 provides a critical review of the research literature relevant to this study. First, this chapter provides a theoretical framework for teacher assessment which describes teacher-based language assessment and its characteristics. The terms consistency and validity in teacher language assessment are clearly defined and discussed. This is followed by the introduction of the concept of assessment bias, and a discussion of the ways in which it is perceived as a threat to consistency and validity in oral language assessment. Next is a critical review of previous studies conducted to investigate the issues of variability in oral language assessment. Critical reviews of the theoretical background to teacher assessment decision-making and its components are presented. This is followed by an analysis of the research gaps and the development of a proposed conceptual framework for teacher language assessment decisions.

The review of literature in this chapter focuses on both variability in teacher assessment and teacher assessment decision-making in language instruction. However, little has been documented in this latter area of research in relation to EAL/D classroom instruction. Therefore, due to the paucity in the literature, a number of studies reviewed here are from large-scale language assessment and teacher general decision-making. Also, in some cases, some research areas have not yet been investigated, resulting in thin evidence of literature. Thus, a review of studies in closely related research areas was included.

2.2. Teacher-Based Assessment: Why, What and How

Before presenting the purposes, content and forms of teacher-based assessment, it is necessary to understand how the term is variously defined. One definition coined by Angelo and Cross (1993) is teacher-based assessment is ‘an approach designed to help teachers find out what students are learning in the classroom and how well they are learning it’ (p.4). Accordingly, the term is characterised to be learner-centred, teacher-directed, mutually beneficial, formative, content specific and ongoing. If we view teaching as a process of decision-making (Shavelson, 1973), teacher assessment decision-making is the process of gathering information about students that can be used to aid teachers in decision-making process (Anderson, 2003). Although teacher-based assessment may be defined in several ways, the nature and purpose of teacher-based assessment should remain the same.

It is argued that teacher-based assessment is important for the entire learning process.(Stiggins et al., 2004). In classroom settings, teachers assess students for many reasons. First, they are usually required to report on student achievement to the principal, parents and other stakeholders (Anderson, 2003). In addition, teacher assessment is used for improving learning (Angelo & Cross, 1993). In this case, assessment results are used as feedback (Black & Wiliam, 1998b; Hattie & Timperley, 2007) to show students their strengths and weaknesses, and to help guide them as to what they should do to improve their performance. This is also known as assessment for learning (Berry, 2008; Black, 1986; Black, 2004; Stiggins, 2002). Apart from utilising assessment results in classrooms to evaluate and promote learning, teacher assessments are also used concurrently for evaluating and enhancing instructional practices (Shepard &, 2001). The assessment of students’ understanding is used as a basis to adjust teaching plans, curriculum content and instructional strategies (p. 67).

The particular question that arises with teacher assessment is the question of what to assess. Teachers usually conduct assessments of their students' work after they complete a section of the curriculum and this can range from a chapter or even a unit of work. In this sense, assessment activities help identify whether students achieve specific learning outcomes. Popham (2014) refers to this as curriculum-driven assessment. Another kind of assessment is decision-orientation or decision-driven assessment. This means that, before creating assessment tasks, teachers must consider *in advance* the kinds of decisions which will be made based on the assessment results. Instruction and assessment must reflect appropriate content, meaning that assessment should be based on some sort of *standard*. As noted by Popham (2014), the 'standard' for classroom assessment is of two types - *content standards* and *achievement standards*, and teachers should carefully consider which to rely on to assess their students.

Another concern is how the *how to assess* question is addressed. Determining how teacher assessment is conducted involves selecting an assessment approach and particular assessment item types. The first decision to be made is between the two most commonly used assessment approaches e.g., *norm-referenced* and *criteria-referenced* (Bond, 1996; Knight, 2001; McNamara, 2000; Popham, 2014). A *norm-referenced* approach describes a scenario in which a student performance is interpreted and assessed in comparison with previous performances on the same assessment task(s) by a group of students known as the norm group or with performances of the students in the same group. In contrary, *criterion-referenced* approach assesses student performance using criteria developed from curricular aim. The performances are interpreted 'according to the degree to which the curricular aim has been mastered' (Popham, 2014, p. 60). Another answer to the how to assess question is also the selection between selected-response and constructed-response assessment schemes. As their names infer,

selected-response scheme is a scenario by which students respond to assessment tasks by selecting given options (i.e., choices), and constructed-response scheme means students must construct or produce their own responses to the tasks. The selection of assessment scheme also helps guide what types of task to be used.

2.3. A Shift in Assessment Theory

In the past, in search for a theoretical rationale for classroom assessment, researchers and assessors borrowed theories originally developed to improve validity and reliability in psychometric assessment. However, some educational researchers (Anderson, 1998; Brookhart, 2003) argue that theoretical frameworks developed for large-scale testing do not fit the practice of classroom assessment, because these two cultures of assessment are significantly differentiated from each other in assessment purposes, procedure and use of results. Therefore, over the past decades several alternative frameworks theorising classroom assessment practice based on constructivism (Bruner, 1986; Piaget, 1970; Vygotsky, 1980) have been developed. Of these, particularly relevant to this thesis are models proposed by Anderson (1998) and Brookhart (2003). These models will be examined in turn in the sections below.

2.3.1. Anderson's framework for alternative assessment

By contrasting with the traditional perspective of assessment and searching for a more suitable theoretical framework grounding school-based or classroom assessment practices and research. Anderson (1998) proposes a theoretical framework from an alternative perspective. Grounded in constructivism theory, Anderson's model of assessment presents a set of theoretical and psychological assumptions about alternative assessment. The first difference between the two perspectives is posited in the

theoretical assumption about knowledge. From a psychometric perspective, knowledge is consensual, with the same meaning for all individuals everywhere (Berlak, 1992), so reaching a meaning consensus among people is likely to be a reality. However, if it is assumed that knowledge has multiple meanings (Roderick, 1991), the assumption that everyone can reach a consensus about meaning is impossible, as everyone perceives and understands a piece of knowledge in their own way and their perception and understanding are different in different contexts.

The second difference between the two perspectives is in the view of learning. In the traditional view of assessment, learning is treated as a *passive process*. Students are described as an ‘empty vessel’ and teachers’ job is to ‘‘fill’’ the students by making deposits of information which [the instructor] considers to constitute true knowledge’ (Freire, 2000). The emphasis is not placed on learning how to do things but on learning about things instead (Anderson, 1998). However, from the alternative perspective, learning is regarded as an active process through which students alter their understanding by looking for new meanings (Greene, 1988); and learning occurs when students produce knowledge, rather than reproducing knowledge (Newmann & Archbald, 1992).

In traditional assessment practices, assessment and instruction have traditionally been described as separate components, and assessment takes priority after instruction (Bintz, 1991). Students are assessed by some type of test and the results are assumed to be representative of their performance. It is noted that ‘the test functions as a ‘‘dipstick’’ into the ‘‘oil tank’’ of a student’s achievement’ (Brookhart, 2003, p. 8). By contrast, in alternative assessment practices, process is equally evaluated as product. Specifically, what and how a student learns are taken into consideration (Johnston, 1992).

In terms of the purposes of assessment, psychometric assessment is assumed to document learning. From this perspective, assessment is to monitor learning, and students are assessed, categorised and ranked. In other words, students who do not know are judged by comparing their performance with that of students who know (Anderson, 1998). Alternatively, since instruction is all about helping students to succeed, facilitating learning must be the primary purpose of assessment (Johnston, 1989; Wolf, 1992). Assessment results function as constructing feedback to students about their learning. Feedback reflects their strengths and weaknesses, enabling them to self-direct and adjust, where necessary, to gain progress in their learning (Black & Wiliam, 1998b; Gipps, 2012).

In addition, to ensure assessment validity, instruction and assessment need to be integrated (Anderson, 1998; Brookhart, 2003). From the traditional perspective, assessment is considered to be objective, while in the alternative assessment culture it is considered subjective. Conversely, in the psychometric culture, although assessment is viewed as objective as assessment is separated from instruction, such separation may threaten the validity of assessment. Tests driven by psychometric approaches that make snapshot assessments may not measure what is instructed. At the other end of the scale, assessment in the alternative culture is challenged by its subjectivity (Bintz & Harste, 1994) because teachers teach and assess. However, in terms of validity, assessment does not make sense without taking the process of learning into consideration, as a student cannot be better assessed by a stranger (Wiliam, 2001).

In terms of power and control, assessments grounded on traditional assessment theories reflect a hierarchical model of power and control (Anderson, 1998). In particular, the teacher is the only person who has the power to decide the content of teaching and assessment (Sessions, 1995) without considering student participation in

making such decisions. An alternative perspective assumes that the power to make decisions about what to be learned and what to be tested should be equally shared by both the teacher and their students (Anderson, 1998).

In conclusion, building on a constructivist perspective, Anderson places emphasis on 'a more democratic stance' in the assessment process. Accordingly, students are given a great deal of power in making decisions about what they have learned and on how they are assessed, developing rating criteria and conducting peer and self-evaluation (Anderson, 1998).

2.3.2. Brookhart's classroometric perspective of assessment

One issue of concern drawing much research attention is the quality of classroom assessment and how this is characterised through the use of the concepts of reliability and validity (Brookhart, 2003). In the absence of conceptual and theoretical frameworks to shape classroom assessment, these concepts (originally created and used for large-scale measurement contexts) have been recently adopted, although such adoption does not align well with classroom contexts (Brookhart, 2003; McMillan, 2003; Moss 2003; Smith 2003). Principles and standards retrieved from psychometric tradition seems irrelevant for teacher assessment (Leung, 2005). The use of traditional measurement concepts, communicative learning approaches and poorly satisfied demands for feedback among teachers and learners may have resulted in misleading interpretations of assessment results of student learning. Therefore, there has been a call for a shift in assessment theory, from psychometric (or traditional theory in relation to large-scale assessment) to a more relevant theory for classroom assessment. Brookhart (2003) proposes a theoretical framework from another alternative perspective called

‘classroometric’ that conceptualises teacher assessment as a process with three different but interrelated components.

2.3.2.1. The nature of the relationship between the measure and the measured

As defined by (Brookhart, 2003), the measure is a test or an assessment task, while the students are the measured. From a psychometric perspective, the test is ‘external to inferences made and actions taken’. As mentioned earlier, the test functions as a dipstick into the tank of a student’s achievement. The test observes the student, collects information about their achievement, and inferences are made based on the collected information without considering how the instruction is carried out, how learning occurs or the student’s perception of assessment. The validity goal of the dipstick is to make a meaningful inference about the student and to ensure that the assessment information is effectively used. Conversely, from a classroometric perspective, inferences made and actions taken based on assessment results are internal to the measurement process (Brookhart, 2003). It is highlighted that teachers make an assessment that becomes part of the students’ learning process and psychology. Hence, inferences and actions based on such assessments make immediate and internal changes in the measured students (Moss, 2003). Together with teachers, students are joint observers who make inferences and take actions based on the assessment information in a formative environment. Students are aware of what the assessment information is about and how it functions to facilitate their learning. In addition, a classroometric perspective views the primary aims of assessment as checking a student’s accomplishment of certain learning outcomes and goals and facilitating learning and grading. Further, assessment also assists teaching on a daily basis. This can be clearly seen in the way in which assessment information helps teachers to evaluate the

effectiveness and authenticity of their instructional design and the strategies on which appropriate actions are taken or adjustments are made to improve teaching. It is also argued that the teacher is internal to the assessment process, as feedback by the teacher is part of the assessment information (Black & Wiliam, 1998a). From a classroometric view, the validity goal of assessment is ‘understanding the role of that assessment information ... in the ongoing classroom learning environment’ (Brookhart, 2003, p. 9). It is understood that a student’s achievement is assessed by comparing it with the ‘ideal’ work previously defined in the learning outcomes. As Moss (2003) suggests, the teacher’s role in the classroom is to interpret students’ capabilities at any stage of their learning process to assist their learning and support.

2.3.2.2. Construct-relevant and construct-irrelevant variance

The purpose of any piece of assessment is to measure a specified construct. As Gipps (2012) noted, the construct is embedded in, and clearly and specifically defined by, a conceptual framework that clarifies the relationship between assessment scores and the construct. In large-scale assessment, the content specifications are descriptions of a domain that are the learning outcomes or objectives. The measurement contexts are construct-irrelevant and considered ‘an extraneous variable which have to be managed and neutralised’ (Davison & Leung, 2009). Therefore, the test must be administered in a standardised way (Brookhart, 2003; Davison, 2007), implying that the equation of scores across contexts and assessment forms is a reality. What are viewed by a psychometric perspective as the greatest weaknesses are the inherent strengths of classroometric assessment (Davison & Leung, 2009). From a classroometric perspective, the context of assessment becomes construct-relevant. As Brookhart (2003, p.7) notes, one of the ways to consider classroom assessment is the integration of

instruction and assessment—the content of assessment tasks depends a great deal on what is instructed in a classroom. Thus, the same task can be perceived differently in two classroom contexts. The test or assignment tasks are part of instruction and a sound assessment is an ‘episode of genuine learning’ (Wolf, 1992). A test or an assignment is framed to help students’ learning. Further, the content specification of assessment describes not only learning teaching objectives but also teaching modes. What teachers believe, how they conduct instruction and what they understand about the subject matter and students should be issues of concern in terms of validity (Brookhart, 2003, p. 10). In addition, classroom assessment is internal to the process of teaching and learning as it is developed and conducted by teachers for different purposes. Therefore, the criteria to evaluate validity should take into account whether assessment has any contribution in part to teaching and learning (Moss, 2003).

2.3.2.3. *Reliability and error*

Under a psychometric lens, reliability equals consistency (Popham, 2014) and reflects stability over irrelevant factors such as occasions, time, tasks, raters and so forth that are treated as facets of ‘error variance’ (Brookhart, 2003). This is to consistently rank students on a scoring scale for norm-referenced scoring or stable categorisation of students along an achievement continuum for criterion-referenced scoring. By defining reliability in this way, it is commonly stated that reliability in the classroom is not necessarily important, as today’s judgement errors can be compensated the following day (Shepard, 2001). However, Brookhart argues that judgement errors made on one day reflect instructional decisions for that day, and ‘an instructional opportunity’ is missed and cannot be added the following day. Consequently, this definition of reliability is solely for large-scale assessment, not for classroom assessment. The

reliability goal of classroom assessment is to stably predict the gap between students' actual work and what they are expected to achieve. Assessment is not to rank students on a scoring scale. It is especially important that assessment can reliably categorise students on a development continuum of work quality and, thus, provide accurate diagnostic information about their weaknesses.

The traditional view of validity is to do with the adequacy of a test or an assessment activity in relation to the measurement coverage. In other words, the content covered by the test is able to represent a larger body of content that the test aims to measure (Akbari, 2012). However, it is often challenging to decide the content, and in most cases, underrepresentation or inadequacy of what should be assessed is often seen (Messick, 1996). In addition, validity is also used to refer to the extent to which a new test correlates with existing tests of the same skill to see the degree of correlations between the test takers' scores (Akbari, 2012). This is referred to as criterion-related validity. The concept of validity is even more complex when "recent approaches view validity to be concerned with the inferences that are made of test scores" (Akbari, 2012, p. 33). However, it is not in the scope of this study to engage in this debate; what this study is primarily concerned with is reliability, or in the current parlance, trustworthiness.

Reliability is the consistency of the process of assessment, meaning that the results of a test should remain stable over times or in different conditions. McNamara (1996) argues that teachers are an important factor in enabling validity and reliability in assessment. Traditionally, reliability is seen as dependent on validity. However, the reliability of a test does not guarantee that the test is valid. Other researchers (e.g. Davison & Leung, 2009) have argued that in teacher-based assessment validity and reliability are so intertwined and context-dependant that the term 'trustworthiness' is a more appropriate

term to capture the complexity of the relationship between validity and reliability. As defined by the TEAL project, “trustworthiness means honest, valid and reliable assessments that really do assess what they claim to, and assessment procedures that produce consistent results, when administered in similar circumstances, at different times and involving different raters”. It is in the scope and of interest of this study to examine trustworthiness of teachers’ language assessment.

Overall, classroom assessment should provide students, parents and teachers with most of the information about their learning, including their strengths and weaknesses directly associated with the subject matter, their behaviour and plans for future improvement (Brookhart, 2003). Theoretical concepts about validity and reliability have underpinned most of the recent discussion about classroom assessment quality. The past decade has witnessed a transformation in assessment theories, in response to an alarming level of inappropriateness in applying theories that were originally developed for large-scale testing in classroom assessment (see Anderson, 1998; Brookhart, 2003; Moss, 2003; Smith, 2003). Students are no longer considered a subject of assessment—they proactively participate in the entire process of assessment. They participate in what is to be taught and tested (Anderson, 1998); and they are the primary users of their assessment results in that they are aware of their strengths and weaknesses and can make further plans for more effective learning (Brookhart, 2003).

Theoretical frameworks developed from alternative perspectives are important to, and indeed improve upon, teacher assessment practice and, thus, support student learning. Greater importance is attached to the role of such frameworks in improvement of practice and in taking account of the inherent variability of teacher assessment.

2.4. Variability in Teacher-Mediated Language Assessment

Teacher-mediated language assessment or teacher-based assessment—using humans as assessors in language tests or assessment activities—is usually seen as a problem of reliability (McNamara, 2000). Reliability equals consistency (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.), 1999; Popham, 2014). Consistency in educational assessment and in language assessment is referred to as the degree of stability of individuals and groups in their behaviour. Examples of identical behaviour of the same person in all pertinent aspects are rarely found (American Educational Research Association et al., 1999). Consistency is categorised in three different ways (Popham, 2014). The first kind is *stability consistency*. This is the extent to which the assessment results are stable, even when assessments are made at different times (Luoma, 2004; McNamara, 1996; Popham, 2014). The second kind is *alternate-form consistency*. This type of consistency refers to cases in which two forms of the same assessment instrument are equivalent and apparently measure the same thing. The third type of consistency is *internal consistency*. While stability consistency and alternate-form consistency deal with the number of assessment administrations, internal consistency deals with the degree to which items of a single assessment instrument function properly. Thus, internal consistency indicates whether assessment items are doing their job of measuring effectively.

Consistency, in regard to teachers as assessors, is the ability to give accurate and stable assessment (Luoma, 2004; Taylor, 2006) and consistency implies that the assessment results are dependable for decision-making (Luoma, 2004). However, in teacher-mediated language assessment, consistency is challenged by an inherent

characteristic of teacher assessors, that is, teachers' subjectivity. Therefore, subjectivity is the problem that has to be acknowledged and managed (McNamara, 2000).

Management of assessors' subjectivity is a complex process involving the assessors themselves and their inherent variables (McNamara, 1996) and the subjectivity of teachers is reflected in their variability when they make assessments. Emphasising the concern of examining teacher variability in language assessment, Lado (1961) noted that in selected response or objective tests assessor variability was actually nil and; therefore, was not necessarily a factor. However, in constructed-response tests such as essays or speaking tests, assessors' variability can be considered a major factor of inconsistency in assessment.

Variability in teachers' assessment is presented in several forms. To illustrate, as suggested by McNamara (1996), similar to raters in a psychometric perspective of language assessment, some teachers are more stringent or more lenient than others. This is also known as inter-rater (teacher) consistency. Alternatively, teacher consistency is also demonstrated in the way in which teachers interact with their students. Specifically, one teacher may show severity towards one group of students but show leniency towards another group. In fact, research (Eckes, 2005; O'Loughlin, 2002) has found that some assessors are more favourable towards the performance of students of one gender than the performance of students of the other gender. In addition, in oral language assessment, when students come from a certain group of accents, ethnicity and so forth this may also affect consistency in teacher assessment. Another alternative form of teacher consistency is reflected in the way in which teachers judge students' performances on several types of tasks and genres. Through assessing performances generated from tasks with these two characteristics, teachers' unknown and unexpected assessment behaviours may be elicited (Kim, 2009). Teachers' consistency is also

presented in their interpretation of the assessment rubrics. For example, one teacher may consider content important in oral communication and; therefore, may give especially stringent assessment in this category and more lenient assessment in other categories. Conversely, the opposite may be true for another teacher. The final alternative consistency—internal consistency—requires a teacher to be stable or to give stable assessment on separate occasions.

From a psychometric perspective, variability is considered problematic and has a negative effect on assessment reliability. Variability is also referred to and understood as assessment bias in which ‘qualities of an assessment instrument that offend or penalise a group of students because of students’ gender, race, ethnicity, socioeconomic status’ (Popham, 2014, p. 127) in the test content and also in the scoring process. Offensiveness and unfair penalisation are two forms of assessment bias in content. Offensiveness is the extent to which ‘negative stereotypes of a certain subgroup’ are found in the test content. Unfair penalisation occurs, in the absence of offensiveness, when the content of a test is disadvantageous to one subgroup, but beneficial to another. In the scoring process, variability is displayed in the form of teacher bias and is referred to as unexpected interaction between raters’ assessment and students’ performances or other factors (Schaefer, 2008). From a psychometric assessment viewpoint, there is no room for bias. In fact, it is argued that, apart from validity and reliability, another criterion to evaluate the quality of educational tests is ‘absence-of-bias’ (Popham, 2014). Thus, if any assessment bias is detected, regardless of test content or scoring procedure, the test scores are distorted and; therefore, unreliable. To prevent assessment bias or variability in the scoring process when developing large-scale standardised tests, testing institutions try to include objective items such as multiple-choice questions or short-answer questions. Further, due to awareness that subjectivity is an inherent variable of

human assessors, rater-mediated language assessment was rejected from the 1950s to the 1960s. The first automated essay scoring system by Page (1966) was introduced as an alternative to human raters. Since then, more automated essay scoring systems have been generated such as a new version of the Project Essay Grade (Page, 2003), the Intelligent Essay Assessor (Landauer, Foltz, & Laham, 1998), e-rater (Burstein et al., 1998) and IntelliMetric (Elliot, 2003). However, although a great deal of research has been done to investigate the reliability of e-raters in comparison with human raters and no significant differences have been found, much debate remains over the reliability of e-raters as a substitute for human raters, even in large-scale testing operations.

Conversely, variability as inherent to humans is seen by the alternative perspective a beneficial and integral to assessment. Since subjectivity is inherent (McNamara, 2000) and the primary purpose of classroom assessment is to help make improvement, if communicated to teachers, information on variability can be treated as a source of valuable feedback. Once teachers are aware of their biased behaviour, this awareness to some extent helps teachers monitor their performance and behaviour, enabling adjustments and enhancements to be made. Further, variability among teachers in assessment is fundamental, as they are supposed to hold different perceptions of students' achievements, to have a different understanding and interpretation of assessment criteria and, thus, make different assessment decisions. As Davison and Leung (2009) note, 'it is not necessary to have complete consensus; that is, teachers do not need to agree to give identical scores; some variation within the range is expected' (p.409). To make fair and accurate decisions about student achievement, Hamp-Lyons (2009) proposes teachers collaborate in making assessment decisions within each school and across schools.. At the same time, variability can be made visible and (re)negotiated through such professional conversations. The trustworthiness of assessment in

classroom situations focuses on the process by which teachers show their disagreement and justify their points of view rather than on complete agreement (Davison, 2004).

2.5. Previous Studies on Teacher Assessment Variability

The aim of this study is to examine differences in teachers' assessments of oral English communication skills and to explore the factors that drive their decision-making behaviour and how they assess. Given that teacher-mediated assessment is predicated on the involvement of a human assessor, also known as an agent of subjectivity, and that little is known about this line of research in classroom settings, a great deal of literature on psychometric language assessment is reviewed as a relevant source of reference in this study. As stated earlier, examining variability in teachers' assessment should include exploring as many factors that influence their assessment consistency as possible. However, this PhD study focuses only on the influential factors associated with teachers, students and tasks. Furthermore, given that both oral and written language assessment practices involve teachers making informed decisions about language learners' work, this study, wherever relevant, also reviews research studies in large-scale language assessment contexts.

2.5.1. Variability related to teachers' experience

One of the influential factors in the consistency of teacher assessment is the nature and extent of teachers' assessment experience and this has been extensively researched. A typical example is Barkaoui (2011), who undertook a study to identify the effects of rating methods and assessors' experience, Barkaoui took a set of sampled ESL compositions rated by novice assessors ($n = 31$) and experienced assessors ($n = 29$). The analysis of multi-facet Rasch on the two groups revealed that novice assessors

and experienced assessors were significantly different from each other in overall severity, inter-rater consistency and internal consistency. In terms of severity, as a group, experienced assessors were consistently more stringent than novice assessors and experienced assessors were shown to be consistent with their group. In addition, analysis of holistic and analytical scorings showed that novice assessors were consistently more lenient than experienced assessors. The previous work of Barkaoui (2010a, 2010b) explain how experience affects holistic scorings and the researcher also reported assessor differences due to experience. Specifically, inexperienced assessors highly valued students' argumentation in essays and their assessments indicated more variations, whereas experienced assessors tended to be more stringent and more highly valued linguistic accuracy. These findings indicated that assessors' experience played a significant role in rater-mediated language assessment.

Another study to investigate the effects of assessors' experience was conducted by Leckie and Baird (2011) in a large-scale operational testing setting in England. A total of 34,920 scores assigned by 689 assessors including 135 very experienced team leaders, 372 experienced assessors and 182 new assessors were analysed. The results showed that in terms of severity all three groups of raters performed differently. However, the observed differences were not significant, meaning that rater experience did not necessarily determine whether they were lenient or stringent. In terms of absolute agreement in scoring, in comparison with the consensus scores rated by expert assessors, scores assigned by the most experienced groups were significantly different while experienced assessors' scores and new assessors' scores were not.

Cumming (1990) also conducted a study to investigate whether there were differences between novice teachers and experienced teachers in assessments of ESL writing performance. In this study, 13 volunteer teachers including seven student

teachers with no teaching experience and six ESL expert teachers were asked to holistically assess 12 compositions at intermediate and advanced levels selected from a pool of 147 papers in an ESL placement test. From the analysis of the teachers' verbal report, Cumming found that teachers used 28 decision-making behaviours to interpret and assess students' performance. Both teacher groups used a comparable number of behaviours. However, they were qualitatively different in the way in which they controlled their assessing behaviour and in how much they focused on composition aspects. Specifically, expert teachers were more flexible in their ratings. This result was explained by the fact that more experienced teachers usually employed their own assessment categories which were not present in the assessing scale (Eckes, 2008; Wolfe, Kao, & Ranney, 1998). They counted the number of ideas in each composition more frequently to decide the overall performance. Conversely, most of the novice teachers' behaviour involved 'editing phrases' or correcting errors.

Showing the same tendency, in an investigation to examine the effects of training on raters' performance, Weigle (1998) asked eight inexperienced assessors and eight experienced assessors to score 60 essays twice, before and after training, in the context of an English as a second language placement examination. Using FACETS to analyse scores, in terms of severity and consistency, Weigle found that, before training, novice assessors were more stringent and less consistent than experienced assessors. After training, the differences between the two groups of assessors were significantly reduced in consistency and slightly reduced in severity. New assessors were reported to display more variability than experienced assessors. This implies that, with or without training, rater experience has a considerable effect on assessor performance in second language assessment. In a further study, Weigle (1999) also observes the effects of rater experience and reports that inexperienced raters continue to give more stringent ratings

at first and more lenient ratings later, whereas experienced raters remain stable in terms of severity.

The implications of this research for my study are that the degree of variability in teachers' assessments may be constituted by their experience. In large-scale assessment contexts, training is suggested as an effective resolution to the problem; and if this persists, even after training, teachers with high degrees of variability may be eliminated from the operational assessment process. However, this study is to explore how experience affects teachers' assessment decisions.

2.5.2. Variability related to teachers' and students' gender

In oral language assessment, assessors are usually viewed as a source of variability (Bachman, Lynch, & Mason, 1995; Davison & Leung, 2009; Luoma, 2004; McNamara, 2000) and gender-related bias can be observed in association with the gender of students, interviewers and assessors (Brown & McNamara, 2004). Nevertheless, there is research that investigates gender effects for both assessors and students. Several studies have been conducted to examine gender effect on language assessment.

For example, a study was conducted by Eckes (2005) to investigate the effects of raters in writing and speaking sub-tests of the Test of German as a Foreign Language. This large-scale study involved 1,359 student participants including 747 females and 612 males for the writing section, and 1,348 participants including 741 females and 607 males for the speaking section. A multi-facet Rasch measurement approach was mainly employed for data analysis. The results show that, on average, assessors tended to be more lenient with female examinees than with male examinees. Although there was

insufficient evidence to conclude that assessors displayed gender bias, in some individual cases assessors were willing to give higher scores to females to males.

To complicate the overall picture of gender effect, Carroll (1991) conducted a study to investigate the effects of gender in an interview test and his findings showed that male assessors were more lenient than female assessors and male examinees received higher scores than female examinees from both male and female assessors. Overall, both male and female assessors gave higher scores to male candidates. Carroll's findings were opposite to the findings of Porter and Hang (1991). Porter and Hang found that female assessors tended to be more lenient when assessing the performance of examinees from different nationalities. Despite apparently not reporting on the effects interviewers' gender had on scoring, the studies of Carroll (1991) and Porter and Hang (1991) suggest assessors' and candidates' gender had significant effects on assessments.

Lumley and O'Sullivan (2005) conducted research to test the hypothesis that variables such as the task topic and the gender of the topic presenter and of examinees had significant effects on the assessments of speaking tests. To eliminate the interlocutor variable that was used in previous studies, the researchers had the test administered in a language laboratory and the performances of all participants ($n = 894$) were audiotaped. The audio recordings were then rated by 30 trained and accredited raters. The results showed that female candidates performed slightly better than male candidates. In addition, the gender of the audience had an effect on the candidates' performance. Specifically, male candidates performed better if the audience was male, whereas female candidates were advantaged when there was a female audience.

Building on his previous work (O'Loughlin, 2000), O'Loughlin (2002) investigated the role of gender in oral interview tests such as IELTS. The data were

derived from audio recordings of the performances of 16 candidates (eight males and eight females). All candidates were interviewed two times, once by a female interviewer and once by a male interviewer. Each of the performances ($n = 36$) were assessed by four raters (two males and two females). In contrast to previous studies, no gender-related effects were found in the assessment of the IELTS interview. First, it was found that the Z-scores were within an accept range of -2 and $+2$, indicating that the assessors were not significantly more stringent to candidates of one gender than the other gender. Second, the gender of the assessors was found not to affect their severity, suggesting that both male and female assessors were consistent in terms of severity. It was also found that male assessors' scores assigned to male candidates were not significantly different from scores they assigned to female candidates and the same was found with female assessors. However, the findings of this study are unusual, and the differential treatment of examinees in language assessment due to gender remains a technical and ethnical concern (Brown & McNamara, 2004).

The implications of this research for my study are that teacher gender may have effects on the way they make judgements about student outputs. This study aims to measure the effect of teacher's gender and student's gender on teacher assessment decisions in classroom contexts.

2.5.3. Variability related to teachers' main exposure to a language group

Although this study did not aim to investigate the effect of teacher language backgrounds on their assessment decision-making, a brief review on this issue is necessary. The reason being is that the effect of teachers' exposure to a language group in assessment is quite new and, thus, has not yet been reported in literature. Further, exposure to a language and teaching or working with students speaking that language to

some extent gives teachers a certain amount of knowledge about the learning characteristics of those students. These characteristics include the strengths and weaknesses, learning styles and so forth that are typical to students from that language background. The more teachers work with students from that language background, the better they understand their students' language learning. The way in which such knowledge and understanding may have an effect on teacher assessments may not be equally comparable to the way in which a teacher's language background affects their assessments. In other words, the effect of the teacher's knowledge and understanding of a language on their assessment may be comparable, though not necessarily the same as, the effect of the teacher's language background on their assessment. Therefore, a review of the effect of teachers' language background on their assessments will shed light on understanding the role of teachers' main exposure to a language group on their assessment judgements.

The number of ESL/EAL/D learners has been constantly increasing in recent years, resulting in a greater number of teachers who speak languages other than English. When working as assessors, those teachers' language backgrounds may be viewed as a salient variable affecting the reliability of assessor-mediated language assessment. A great deal of studies in language testing contexts have been conducted to compare assessments by assessors with different language backgrounds, both first language (L1) and second (L2). Of those studies, some have focused on L1 effects on score assignment (Brown, 1995; Caban, 2003; Johnson & Lim, 2009; Shi, 2001; Xi & Mollaun, 2009; Yan, 2014) and have produced contradictory findings. In some studies, the effects of assessors' language backgrounds on their assessments are proven to be minor or insignificant. For example, Brown (1995) conducted a study to explore how occupational and linguistic backgrounds influenced assessors' performance in the

Japanese Language Test for Tour Guides. Brown found that assessors from different linguistic backgrounds were not significantly different from one another in their assessments. The only major difference between assessors lays in their judgements of specific criteria. Brown then implied that assessors held different perceptions of what they thought was acceptable, and these perceptions appeared constant, regardless of how explicit the rating scale and how standardised the training might be.

Similarly, with the hypothesis that assessors' language background may influence their assessing performance, Johnson and Lim (2009) analysed scores by assessors of the Michigan English Language Assessment Battery (MELAB) of test-takers of languages other than English. The assessors also came from differing language backgrounds including English, Spanish, Korean, Filipino and Chinese. The results showed that the effect of language background on assessor performance was not significant.

Aligned with the findings of Brown (1995) and Johnson and Lim (2009), Xi and Mollaun (2009) found when examining the way in which trained and certified Indian assessors of English judged the oral performance of TOEFL iBT test-takers from different L1 backgrounds there were no differences between assessors from India and operational assessors from the US in score assignment to Indian and non-Indian test-takers. Supporting what has been found by previous researchers, Kachchaf and Solano-Flores (2012) reported similar findings from their research. To examine the effect of assessor language background on the assessment of English language learners' (ELLs) responses to short-answer and open-ended items, the researchers recruited assessors from two different language backgrounds, namely English and Spanish, to score 107 responses of ELLs to a mathematics test written in both English and Spanish. The mean scores assigned by assessors were analysed and generalisability theory was used to

measure the amount of variation. Statistically, Kachcharf and Solano-Flores observed that the difference in the mean scores caused by language backgrounds was significant but relatively minor and negligible. These observations were then explained by the fact that all assessors were accredited bilingual teachers. However, as with other researchers, Kachcharf and Solano-Flores failed to draw a generalised conclusion on the effects of assessor language background when their assessor population was limited ($n = 8$). Further, the task types were simply short-answer and open-ended questions that were not as complicated to score as essay writing and speaking.

Conversely, some studies have claimed that assessors' language backgrounds are a major factor in distorting their assessments. One such study by Shi (2001) investigated the differences between English and Chinese teachers of English when they holistically assessed compositions by Chinese students of English. Although the researcher did not find any major differences in scores assigned by the two groups of teachers, chi-square tests suggested considerable differences in their judgements of rating categories. While English teachers were more generous towards content and language, Chinese teachers tended to be more stringent in organisation- and -length-related categories. Similarly, in a study by Caban (2003) that attempted to explore perception differences of ESL raters of English, ESL assessors of Japanese, English assessors without an ESL background and peer assessors in assessing Japanese medical students, differences were found between the four assessor groups in terms of severity. Specifically, Caban's findings indicated that English assessors were consistently more lenient in pronunciation assessment compared to other groups, while Japanese assessors were found to be consistently more stringent in pronunciation and grammar assessment but more lenient in other rating criteria (i.e., overall intelligibility, compensation techniques and language appropriateness). Despite the rater discrepancies found in her study, Caban was

sceptical about the effect of language background on assessment consistency, but she did not provide a detailed explanation for her scepticism.

Further complicating the issue of language background, Zhang and Elder (2010), in reporting on a study on ESL and EFL teachers of the College English Test–Spoken English Test, found that teachers of English and teachers of Chinese did not differ from each other in their holistic scorings. However, through analysis of their comments it was found that their interpretation and perceptions of test-takers' performances did differentiate. Their comments revealed that Chinese teachers focused more on linguistic forms of test-takers' speech, whereas English teachers paid more attention to communication efficiency. However, the study failed to generalise the effect of language background due to the limited population of teachers and convenience sampling. Further, unguided holistic writing might also be a factor of teacher variability.

Kim (2009) highlighted the fact that most previous studies investigating variability among English speaking assessors and assessors from language backgrounds other than English have been conducted quantitatively as an evidence deficit for any conclusions relating to such variability. Therefore, Kim conducted a study using a mixed method approach to further examine assessors' behaviour. Her findings revealed that both English Canadian assessors ($n = 12$) and Korean assessors ($n = 12$) were found to be consistent. Interestingly, differences between assessors from the two language groups were found in the comments they gave related to the assessment criteria and this was consistent with the work of Zhang and Elder (2010). English Canadian assessors were likely to provide more comments than their Korean counterparts and their comments on test-takers' performance were also more linguistically specific and elaborate.

Another line of research has focused on the effects of raters' L1 background on their severity. An example of this research is a study conducted by Fayer and Krasinski (1987). In comparing the reactions of L1 English speakers and L1 Spanish speakers to speeches produced by Puerto Rican learners of English, the researchers asked the two groups of speakers to listen to the speeches and then complete a questionnaire covering linguistic forms and irritation levels. They found that Spanish speakers were more stringent in rating linguistic forms and appeared more irritated when listening to the speeches compared to the English speakers. However, these findings were contrary to the findings of Kim (2009) who stated that she did not find any dissimilarities in severity between English Canadian assessors and Korean assessors. This suggests that, with respect to severity, Korean assessors could be considered as reliable as their English Canadian counterparts.

This claim about the effect of language background is further supported by a study by Winke, Gass, and Myford (2013). When exploring how assessors' first language affected their assessment of test-takers whose first language was similar to their own, Winke et al. (2013) found that L2 Spanish assessors gave significantly higher scores to test-takers whose first language was Spanish. This was also the case with L2 Chinese assessors of L1 Chinese test-takers. Most recently, reporting on a study using the mixed methods approach, Yan (2014) used the data derived from the Oral English Proficiency Test to assess English proficiency of potential international teaching assistants in the US. The assessors were well trained and came from different linguistic backgrounds such as English, Chinese and Japanese. By triangulating quantitative results and qualitative comments, Yan found that the assessors were satisfactorily consistent, although they did differ from one another in their severity, especially on lower score levels.

The implications of this research for my study are that teachers' assessments may be affected by their first language or by their language backgrounds. It can also be implied that, as in this study, the amount of teacher exposure to students from a language background may shape the way they assess performances by students from a similar language background.

2.5.4. Variability related to teachers' educational backgrounds

The fact that many EAL/D teachers are officially working while they do not hold EAL/D teaching qualifications or are not trained to be EAL/D teachers (Davison & Michell, 2014) is also viewed as a threat to the quality of teacher-based assessment. Teachers' lack of educational qualifications has long been an issue of concern in language education in general and in language assessment in particular. However, little has been reported on this in the literature. In a study to investigate the effects of occupational and linguistic backgrounds on language assessment, Brown (1995) included 33 assessors who were native- and -near-native speakers of Japanese who worked as teachers of Japanese as a foreign language or as tour guides. All the participants were asked to give ratings on the performance of 51 test-takers. The results of this research showed that non-teaching assessors as a group were less consistent and more stringent than assessors from teaching backgrounds. In addition, significant differences were found between teaching and non-teaching assessors in the way in which they used the rating scale. Notably, non-teaching assessors were more willing to assign extremely low scores or high scores to test performances than teaching assessors.

While these studies examined variability in assessors' educational backgrounds by comparing performances of teacher assessors and non-teacher assessors, there has not yet been an investigation into assessments made by teachers with different levels of

educational expertise or qualifications. This is an issue that, as already stated, is becoming more salient with the global spread of EAL/D education.

In terms of professional background or teaching position, in a study to examine how primary and secondary school teachers observe and assess performances by primary school foreign language students, Butler (2009) found that teachers' professional backgrounds had effects on their holistic assessment decisions and attitudes toward observation and significance of criteria. Specifically, what Butler (2009) found was substantial variability among teachers (both within and across the different school levels). In addition, these two groups of teachers were also different in their understanding as well as weighting of different traits of the criteria.

The implications of this research for my study are that, teachers' qualification and teaching position, amongst others, may influence the way they make their judgemental decisions. Variability in teachers' assessments may be caused by how qualified they are in fulfilling their job as classroom assessors and the level of schooling they are working at.

2.5.5. Variability related to students' language background

In ESL classrooms, it is common to hear students speaking English with foreign accents, as they come from different language backgrounds other than English. Therefore, it is likely students' accents influence teachers' perception of students' interactions and, consequently, their assessments. For decades, the effects of accent familiarity on comprehension and communication have been ascertained by speech processing and speech perception researchers (Gass & Varonis, 1984; Tauroza & Luk, 1997). In standardised language tests such as IELTS, the effect of accent on assessment has been a common concern. One such concern was examined in the study by Carey,

Mannell, and Dunn (2011), who aimed to determine how influential accent familiarity was to the assessment of pronunciation. Carey et al. recruited 99 experienced teachers working as IELTS assessors from five different test centres in Hong Kong, Korea and India to voluntarily take part in the research. The results showed a positive correlation between scores and teachers' familiarity of examinees' accents. Specifically, teachers tended to give a high score of 6.0 to speeches with a familiar accent and a low score of 4.0 to an unfamiliar accent.

The issue is further complicated when it is claimed that listeners may hold some stereotypes of foreign accents (Major, Fitzmaurice, Bunta, & Balasubramanian, 2002). Stereotypes are often observed against non-native-accented speech and are constructed in perceptions of both native and non-native speakers of English, by which listening comprehension is distorted. Over decades, researchers have developed a robust pattern of speech evaluation that non-native-accented speakers are normally evaluated less favourably by native speakers of that language. For example, Edwards (1982) found that students' accents had a significant effect on teachers' assessment, with teachers rating accents that were like their own more leniently. Although the study also aimed to investigate other influential factors of teachers' assessment, evidence of accent effects could not be ignored. In another study, Gill (1994) employed 90 standard American English-accented people who were asked to respond to speeches of three different accents such as American, British and Malaysian. The results showed that British- and -Malaysian-accented speeches were less favourably perceived than the American-accented speeches. Introducing a different view into the effects of listeners' stereotypes of accent on assessment is a study on language attitudes by Cargile and Giles (1998). The researchers had standard American English speakers rate Japanese-accented speeches. They found that in terms of status-related traits the standard American English

listeners stringently rated the non-native-accented speech, but more favourably evaluated the speech in terms of attractiveness.

When considering the diversity in cultural and linguistic backgrounds in today's EAL/D classrooms in English speaking countries such as the USA, the United Kingdom, Canada and Australia, the effects of students' accents on assessment is paramount. However, for learning or accountability purposes language assessment should be as close to fairness and reliability as possible, regardless of whatever accent a student might have.

The implications of this research for my study are that where students come from affect how their performances are assessed. To some extent, it is interesting to explore how teachers' main exposure to students from a language group through their regular or everyday work influences their assessment.

In this study, investigating the effects of teachers' exposure to a language group on their assessments is inspired by the well-documented effects of student language backgrounds on such assessments. Student language background is also considered another form of assessment bias. There is a body of research into how students' language background affects teachers' performance as assessors. It is the way in which the language background of students being assessed affects the people who are assessing them. Although these perspectives are slightly different, they are relevant because the EAL/D students from this study were from different language backgrounds learning English as an additional language or dialect. Therefore, a review of the studies that examine student language background effects is important and necessary to provide more foundation for findings on teacher exposure to a certain language group.

2.5.6. Variability related to tasks

Given that assessment in language education is the interaction among assessors, students and assessment tasks (Luoma, 2004; McNamara, 1996), examining the reliability of assessment should address the interrelationships between these three assessment components. A body of research has focused on investigating interactions between assessors and students or students and tasks. Although little has been reported on the effect of tasks on assessors' performance, this issue of concern should not, to any extent, be neglected. The three components of assessment are interrelated to each other; therefore, any generalisations made in the absence of any of them are considered invalid. Among a few studies in this line of research is the study conducted by Kim (2009), who investigated the differences between native and non-native assessors when they made judgements of oral English performances. While the effect of tasks on student performance has been investigated and confirmed (Fulcher & Reiter, 2003), such an effect on assessors was not the focus of previous studies (Fayer & Krasinski, 1987; Galloway, 1980; Hadden, 1991), as researchers only used tasks of one type. Therefore, in Kim's study three tasks (i.e., picture-based, situation-based and topic-based) were included in a test to measure students' diverse oral output. In conducting a bias analysis to search for potential interactions between a group of teachers and tasks, Kim found no negative or positive biases in ratings assigned by teachers of the two groups. Hence, the tasks did not have any influence on raters' assessments.

In Kim's study, the participating teachers were all qualified and experienced and were carefully selected. Therefore, findings that raters were not affected by task types were almost predictable. In some sense, it can be understood that the teachers were trained to rate and the effects of training on raters have so far been confirmed by a large body of research (Lumley & McNamara, 1995; McNamara, 1996; Shohamy, Gordon, &

Kraemer, 1992; Weigle, 1998, 2002). However, effects of task types on assessors from mixed professional backgrounds are worth considering. In addition, potential interaction between assessors and task genres has not yet been reported in literature.

The implications of this research for my study are that there is a tendency that teachers may sometimes assign a score to a student performance depending on a type or genre of the task instead of what the student can actually perform. To further explore the issue, this study aims to examine if there are interactions between teacher assessment decisions and the type of task that students respond to.

2.5.7. Variability related to scoring rubrics

The role of scoring rubrics or ratings scale has been reported by a body of studies in large-scale assessment contexts. For example, in an investigation to examine the roles and effects of rating scale and experience in rating ESL compositions, Barkaoui (2010c) involved 11 novice and 14 experienced teachers rating 12 ESL essays using two different rating scales: holistic and analytical. The results show that teacher assessments tended to be influenced more by rating scale type than by rater experience. As such, when rating holistically raters focused more on student writing while analytical rating made them focus more on the rating scale. This indicates that the rating scale type somewhat shapes the way in which raters' mark. In another study using a mixed method approach to explore the effects of two different rating scales ratings of ESL essays, rating processes and raters' views, Barkaoui (2007) surprisingly found that teachers tended to agree with one another more when they used the holistic rating scale, compared to the multiple-trait rating scale (i.e., analytical). The researcher suggested an explanation for the low reliability of ratings using the multiple-trait scale could be due

to the lack of rater training, specifically in using the new multiple-trait scale. Barkaoui's findings on the outperformance of the holistic scale over the multiple-trait scale is contrary to what was stated by Jonsson and Svingby (2007).

A range of other studies has been conducted to investigate how scoring rubrics affect assessment decisions, in that rating categories had certain effects on ratings. For example, Kondo-Brown (2002) conducted a study to investigate rater bias in rating Japanese second language essays. Data from this study were collected from 234 students and three trained teacher raters. The results show that although teachers were consistent and their scores were highly correlated, they were found to be biased with rating categories of the rating scale. Raters were reported to be lenient on one category but stringent on another. As such, Rater 1 was more stringent on vocabulary but more lenient on content. Rater 2 was stringent on content but lenient on mechanics and rater 3 was stringent on mechanics but lenient on vocabulary. This indicates that teachers did not weight rating categories equally.

In another example, Schaefer (2008) conducted his study to investigate bias patterns of native English raters rating EFL compositions. In his study, Schaefer invited 40 raters to mark 40 compositions by Japanese female students on a written TOFEL adapted topic. The results suggested that while raters rated content or organisation stringently, they were lenient on language use or mechanics. Showing the same tendency, Eckes (2012) conducted a study to explore how rater cognition and rater behaviour are related to each other using a sample of 18 raters from his previous study (Eckes, 2008). The results suggest that raters were different from one another regarding their perceptions of criterion importance. Criteria which were considered important were rated stringently and criteria which were believed to be less important were rated leniently.

The implications of this research for my study are that teachers' assessments tend to be affected by the type of rubrics that are used to assess student outputs. It is also implied that the way assessment rubrics are constructed, whether they are implicit or explicit, may influence decisions of the users – teachers.

Overall, a significant amount of research has been done to understand the nature of variability, especially the factors that contribute to its occurrence. However, to achieve a better understanding of teacher assessment, it is important to gain more insight into the process by which teachers make their decisions.

2.6. Teacher Assessment Decision-Making

Teacher decision-making is a complex process; therefore, it has long been defined in several ways. For example, Hipkins and Robertson (2011) state, 'making decisions about the qualities of specific examples of student work involves the use of a number of different resources' (p.9). . Decision-making is a process with three sequential components: 1) teacher attention drawn to student output, 2) teacher assessment of student output against some given scoring rubric and 3) teacher action or judgement decision (Sadler, 1998). It is also noted that different teachers, at every decision point, tend to refer to and apply different resource types to make their judgement. The way in which teachers come up with different decisions about the same sample of student work depends on details of the sample they attend and the level of attention they pay to the resources.

Another definition of teacher decision-making is one that is proposed by Colton and Sparks-Langer (1993). Accordingly, teacher decision-making is a reflective process with teachers as reflective decision-makers. This reflective process is attributed to the differing characteristics of teachers. One of the characteristics of a reflective teacher is

efficacy—the belief that they are an influential factor of student success. Efficacy motivates teachers and encourages them to ‘look for deeper meanings’ (p. 50).

Flexibility is another attribute of reflective assessment decision-making. This is an important attribute, as teachers are required to view the world through someone else’s perspective. This enables teachers to revisit their perceptions more thoroughly and objectively. Another attribute is social responsibility. This is all about teachers’ care towards social values. They embed this in their decision-making to encourage students to be more socially responsible and to care about others. The last attribute is consciousness, in that teachers are aware of what they are thinking and the decisions they are about to make. Therefore, they can professionally explain and reason their actions.

It is well documented that reflective decision-making is an important prerequisite for effective teaching (Clark & Peterson, 1984; Good & Lavigne, 2017; Wilen et al., 2004). Assessment experts can use knowledge about how assessment decisions are made to ‘reconceptualise assessment principles and suggested practice in ways that further teacher’s goals for students’ (McMillan, 2003, p. 39). This also implies that teachers are required, through training, to obtain crucial knowledge and skills to make instructional decisions. Awareness and understanding of assessment bias allows teacher to understand and improve their consistency and fairness. Furthermore, teachers should also be trained to use assessment to enhance teaching.

Assurance of consistency in teacher assessment decision-making is often challenged by differences in beliefs and practices, which is more likely disadvantageous for some students (Butler, 2009). Thus, Davison (2004) suggests classroom-referenced assessment as an alternative to teacher-based assessment. That reaching a consensus in teacher assessment is impossible does not necessarily mean consistency may not be

improved. Enhancing teachers' mutual understanding of each other's beliefs and practices seems to be, among others, a decisive variable in fulfilling assessment consistency goals.

Assurance of reliability in educational assessment may not be reached without relevant theoretical guidelines. Brookhart (2003) and Anderson (1998) together with many others have created strong conceptual foundations for trustworthy and dependable teacher assessment practice. Similarly, to help teachers make good assessment decisions, they need to be theoretically equipped with conceptual framework that better fits their assessment decision-making practice. Although a range of frameworks have been developed in the past decades, the McMillan (2003) framework has been chosen to be reviewed in the next section due to its close relevance the investigations in this study.

2.6.1. A framework for teachers' classroom assessment decision-making

In line with attempts to better understand how teachers make decisions about student performance and how assessment information is used to support learning and teaching, McMillan (2003) proposes a framework of teachers' classroom assessment decision-making. Details of this framework can be found in Figure 2.1. Accordingly, teachers' classroom assessment decision-making is a process that is constituted by the interplay between: (1) teacher knowledge, beliefs, expectations and values, (2) external factors, (3) classroom realities, (4) decision-making rationale and (5) assessment and grading practice.

The first component of classroom assessment decision-making is internal factors related to teacher knowledge, beliefs, expectations and values. As can be seen from Figure 2.1, there are five sub-categories in this component. One of the teachers' beliefs is that for student success it is important for teachers to do whatever they think is

necessary. This is described as the desire to ‘pull for’ students (see Cizek, Fitzgerald, & Rachor, 1995). The teachers’ philosophy of education is another internal factor in assessment decision-making. When conducting an assessment, teachers tend to rely on their fundamental beliefs about education and highly value greater outcomes for students. The third sub-category is promoting student understanding. Assessment is to gauge student progress, that is, whether students can apply the knowledge and skills they learn to solve problems and make decisions. Teachers want their students to have robust understanding (also see Shepard, 1997) and the ability to generalise and transfer learning. The fifth internal factor of classroom assessment decision-making is varying assessments to accommodate students’ individual differences. Teachers tend to adapt to students’ individual differences by modifying assessments. This is somewhat in line with their belief of ‘pulling for students’ (McMillan, 2003, p.36). Finally, teachers believe that active engagement in learning is imperative to students and; therefore, students must be always motivated to try their best.

External factors that are not under the control of teachers strongly influence teacher assessment decision-making. Specifically, standardised testing places pressure on teachers’ assessment decision-making because high-stakes tests are contrary to teachers’ beliefs and values. Such high-stakes tests are not often used for formative purposes (Shepard, 2000). The pressure in this instance is that both teachers and students are working towards the tests because they are mandated at some key points throughout the schooling process (i.e., years 3, 5, 8 and high schools). This is also true in Australia, where the National Assessment Program—Literacy and Numeracy (NAPLAN) is compulsory and exerts enormous pressure on teachers (Angelo, 2013; Shine, 2015; Thompson, 2014). District policies interrupt teacher decision-making in that they are often in conflict with the values and beliefs teachers develop from

experience. Finally, pressure from parents normally does not have effects on teachers' assessment practices, but may influence grading, which are generally of more concern to parents. Teachers have to avoid conflicts with parents by giving reasonable justification for their grading.

Figure 2.1 also shows that *classroom realities* as a key factor in assessment decision-making, that is, various aspects of the classroom context including social promotion, absenteeism, disruptive behaviour and heterogeneity. Demands created by classroom realities are often contrary to teachers' wants and wills when doing assessments. Further, one of the striking factors influencing teacher assessment decision-making is related to their *decision-making rationale*. Teachers often undertake assessments without an apparent rationale and find it hard to justify how, why and what they do. In some cases, teachers do have a rationale for their assessment decisions, but their rationale is too individualised and does not conform to common assessment principles. Instead, their assessment rationale tends to be a hodgepodge of influences, indicating that the rationale itself is affected by several factors (Brookhart, 1991; McMillan, 2001).

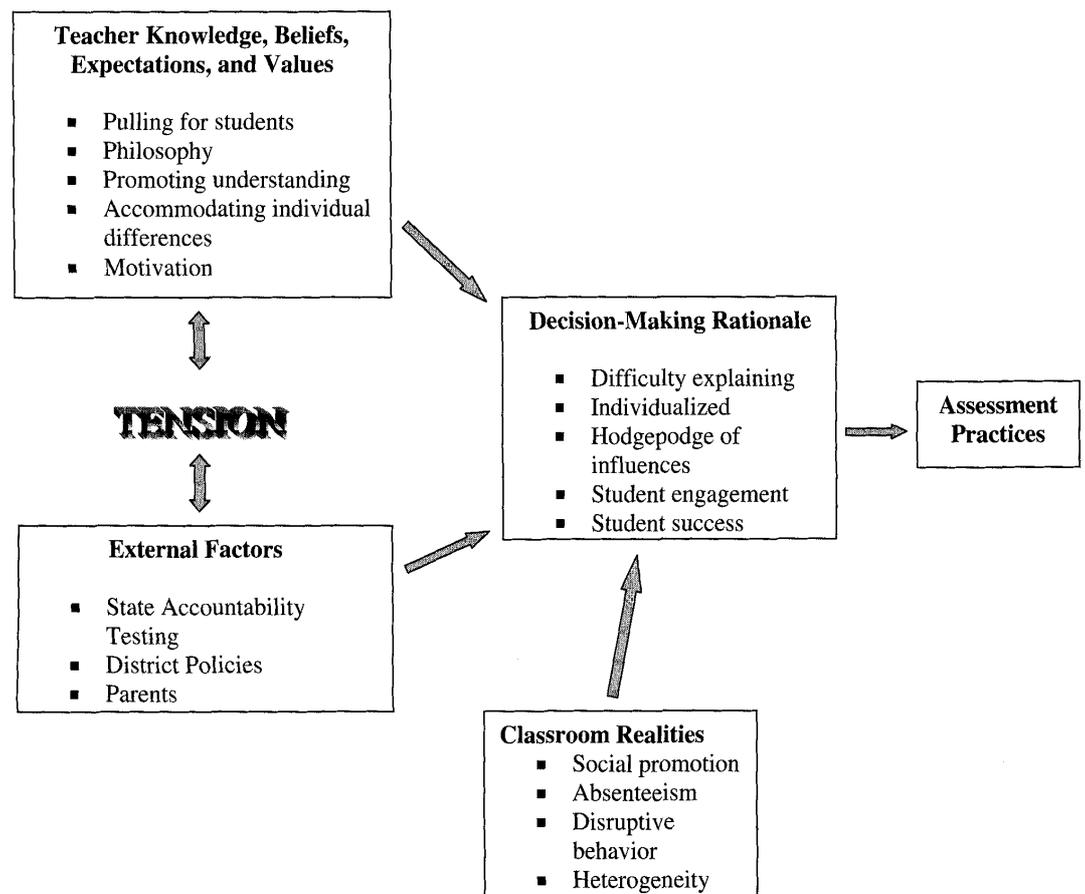


Figure 2.1. Teacher Classroom Assessment Decision-Making.

Source: McMillan, 2003.

This model of classroom assessment decision-making illustrates the relationships between the five components. Notably, the nature of these interactions is that the process of teachers' classroom assessment decision-making is illustrated by the undermining conflict between teachers' internal beliefs and values and external factors. At one end of the continuum of interest, teachers' beliefs and values are often voiced as their philosophy of education. This is consistent with what teachers want to do in assessment and grading practices but is inconsistent with measurement principles. At the other end, pressure from external factors and classroom realities requires teachers undertake assessments in ways that conflict with their philosophy of education. Therefore, accommodating external pressures in assessment decision-making poses real

challenges as well as to teachers as classroom assessors. McMillan's (2003) framework of teachers' classroom assessment decision-making reveals some insights on how teachers make classroom assessments. Moreover, this framework is relevant for assessment in a broader classroom context and can be applied in almost all subject areas. Nevertheless, this framework, to some extent, remains partly relevant and applicable to a more specific area of teacher language assessment (i.e., speaking or writing assessment).

In addition, Rea-Dickins (2001) developed a conceptual framework to support teacher classroom assessment which includes 4 stages including planning, implementation, monitoring and recording and dissemination. However, the researcher also notes "implementation of assessment does not necessarily require a teacher to 'complete' all phases in the cycle. An effective assessment does not need to include all of the above characteristics. What is included, or emphasized, will be dependent on the purpose of the assessment" (p. 434).

The framework proposed by McMillan (2003) suggests that teachers' decision-making in a narrower or more specific area is a complicated process involving several stages., as each teacher needs different theoretical supports to effectively fulfil their duty as assessors. In the following section, a different theory, Gestalt theory, is presented that is believed to provide a better understanding of this component of the teacher assessment decision-making process.

2.6.2. Gestalt in teacher assessment decision-making

Originally developed for better practice in psychology, Gestalt theory takes a holistic view of humans and their behaviours. The notion of Gestalt was first introduced in psychology in the late 1890s by a German psychologist Christian von Ehrenfels (as

cited in Wagemans, Elder, et al., 2012). Later, a more official work was proposed by Wertheimer (1912), who extended it to Gestalt theory and, together with Kurt Koffka and Wolfgang Kohler, founded Gestalt psychology. These Gestalt psychologists were interested in perceiving mind and thought in its totality. To better understand how perception works, Koffka (1935, 2013) introduced Gestalt principles of perception organisation including principles of similarity, prägnanz, proximity, continuity, law of closure and common fate. According to the law of similarity, items that are similar tend to be viewed as a group. For example, as can be seen in Figure 2.2, Image A demonstrates the principle of similarity. Those items that are similar both in visual and auditory stimuli are grouped together. Circles tend to be grouped in colours instead of a collection of fat dots. Image B represents the principle of prägnanz, also referred to as the principle of simplicity, meaning that objects are viewed as simply as possible. In this case, the circles are observed as they overlap one another, rather than a piece of chain. Another principle is proximity. Accordingly, items that are near each other tend to be categorised in one group. As shown in Image C, the circles on the left are observed as a group while those on the right are put in three distinct groups. Continuity is another principle of perception organisation. Image D is perceived as containing two continuous lines that intersect with each other instead of two angles meeting at a point. Image E illustrates the principle of closure. According to this principle, the human brain tends to look for missing information to fill in the gap. When perceiving Image E, we tend to complete the shapes by looking for missing parts to fill the gaps. Finally, Image F is an example of the principle of common fate. That is ‘the tendency for elements that move together to be perceived as a unitary entity’ (Wertheimer, 1923 as cited in Wagemans et al., 2012, p. 1181). Observers perceive objects moving in the same direction as a group.

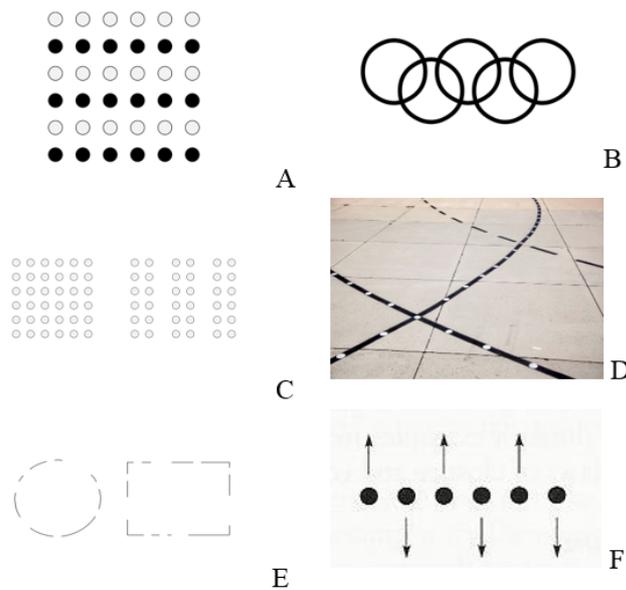


Figure 2.2. Gestalt Principle of Perception Organisation

Source: Cherry (2017) and Connolly et al. (2012)

The primary principle behind the Gestalt laws of perception organisation proposed by Koffka (1922) is that the whole is other than the sum of its parts. There has been misinterpretation that the whole is more or greater than its parts. However, to be correct, the whole is different from the sum of its parts, meaning that the whole should be viewed as the interwoven and meaningful relationships between parts, not simply as addition of parts to make the whole (Koffka, 2013). Representatives of the Berlin School of Gestalt school psychology argue that Gestalt is ‘a whole by itself, not founded on any more elementary objects ... and arose through dynamic physical processes in the brain’ (Wagemans, Elder, et al., 2012, p. 1175). This means that the connection among individual parts is more meaningful than that which the individual parts can do. The behaviour of the whole is not determined by the behaviour of its parts, the intrinsic nature of the whole decides the parts (Wertheimer, 1938). For a clearer understanding of this statement, Heygt, Peterhans, and Baumgartner (1984) use ‘a subjective contour’ to illustrate (see Figure 2.3).

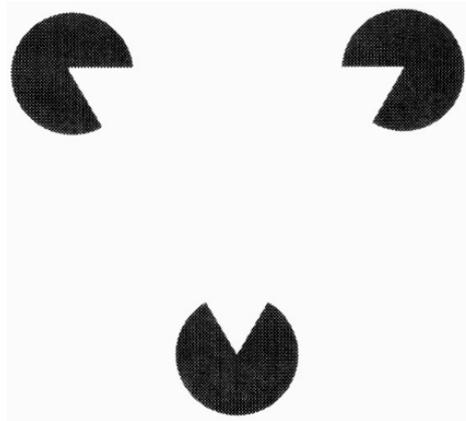


Figure 2.3. A subjective contour.

It can be seen from Figure 2.3 that there are three partial black circles, referred to as parts. There is no real triangle. When taking a closer look, we can see three edges of a triangle in three black circles. The white triangle covers three black circles. It is an independent whole that affects our interpretation of individual elements in the diagram. When we view the circles, white corners with white corners in them, the corners do not appear to be the corners of a triangle. Instead, they look like the Pac-Man character of old video games. When we observe the triangle, the three Pac-Man characters look like three edges of a triangle.

The relations between whole and parts from the Gestalt traditional perspective seem to be difficult for modern readers to perceive. Therefore, more specific operational definitions of such relations are proposed. Specifically, as in a notion called Garner's dimensional integrity, parts, whatever they are, are not perceived independently but holistically (Wagemans, Feldman, et al., 2012). Another notion is emergent features and configural superiority. The former occurs when parts are collapsed into a whole. For example, a collection of seven stripes of assorted colours that are close to each other is viewed as a rainbow that has features that an individual stripe cannot have. The latter indicates that perception of the whole takes place before that of the parts. Another

notion of modern Gestalt psychology is the primacy of holistic properties. Holistic properties are those that cannot be perceived by individual constituents, but their interrelations. This means that holistic properties dominate the constituents when processing information. Overall, the central idea of Gestalt psychology from both traditional perspectives and modern perspectives is the dominance of the whole over its parts in perceptual processing.

Therefore, it is noted that the whole not only is other than the sum of its parts, but also perception and interpretation of its parts. This is a prompt for criticism of Gestalt theory. A concern about the Gestalt theory of perception is this theory has its primary focus on the final perceived product but ignores the importance of insights into how that product is generated (Bruce, Green, & Georgeson, 1996). This suggests that Gestalt theory, to some extent, sheds light on impressionistic or holistic scoring. As such, in impressionistic scoring, teachers view a student's performance as the whole and, thus, assign a single score to decide on that student's ability. However, that single score does not reflect the student's ability in different language dimensions (Weigle, 2002). Therefore, it is not right to generalise student ability by solely basing it on that single score, as this indicates a lack of evidence for making judgements (Thomas, 1994).

The primary purpose of this study is to investigate the process of teacher assessment decision-making to provide better understanding of this process and improve teacher assessment practice. Teachers' perceptions of student works are described in this study as assessment gestalt—the first overall impression and perception of student proficiency and where the students are to be placed in the assessment scale. Assessment gestalt plays a crucial role in the process when teachers make their judgement decisions. Those roles are varied in terms of importance in different assessment styles. Therefore,

Gestalt theory can help reveal what cognitively happens inside the back box - teacher decision-making - and is more theoretically relevant to this study than any other theories. Ultimately, this study aims to help improve the quality of teacher assessment decision-making and this is a socio-cognitive process involving several stages. A socio-cognitive process is a general term “used to describe cognitive processes related to the perception, understanding, and implementation of linguistic, auditory, visual, and physical cues that communicate emotional and interpersonal information” (Suchy and Holdnack, 2013). To better understand what and how, teachers at each stage require relevant theoretical background and support. Like decision-making processes in different disciplines, a decision-maker, in solving a problem, first needs to perceive the problem, then gather information, process information and finally make decisions. Gestalt theory may fit in the first stage by conceptualising teachers’ socio-cognitive processes when they initially perceive students’ work. To better understand the entire process of teacher language assessment decision-making requires the development of a contextual framework that is relevant with the classroom assessment contexts in which all decisions are made for the improved achievement of student learning.

2.6.3. Flexibility and moderation in teacher decision-making

Terminologically, *flexibility* may cause confusion among readers when the same term is used in material science or behavioural psychology to reflect ‘how a person: (1) adapts to fluctuating situational demands, (2) reconfigures mental resources, (3) shifts perspective, and (4) balances competing desires, needs, and life domains’ (Kashdan & Rottenberg, 2010). In classroom assessment, flexibility is referred to as a socio-cognitive process in which teachers contrast their initial thoughts about a student’s work with their beliefs and values and what they have personally and professionally

experienced. This term refers to the likelihood that teachers are open to changing their minds. Flexibility is alternatively referred to as reflection in this study. Flexibility is also referred to as an instrument that enables teachers to validate their initial thoughts about students' work. Flexibility is important to teachers' assessment decision-making; however, little has been documented about this in literature. Colton and Sparks-Langer (1993) suggest a conceptual framework to guide the development of teacher reflection and decision-making.

As can be seen in Figure 2.4, flexibility or reflection in decision-making is a process by which different cognitive, critical and personal elements are integrated. The process starts with influences of the professional knowledge base on interpretations of student works and standards. In making decisions, teachers tend to reflect on their professional knowledge including content, students, pedagogy, context, prior knowledge, personal views and values and scripts. Specifically, as also suggested by Shulman (1987), teachers consider their profound understanding of the subject matter and curriculum (*content*), and student cultural backgrounds and learning styles. Such understanding helps teachers with their choice of *pedagogical* approach (i.e., generic methods and theories as well as pedagogical content knowledge). Further, teachers also consider the assessment *context* in their reflection. Context is important, because the nature of assessment is context-relevant (Brookhart, 2003; Shulman, 1987). Further, teachers' *prior experiences* and *personal views and social values* are strikingly significant to teacher reflection. Reflective teachers tend to relate the present situation and student work to their prior experiences before they take major actions (Kennedy, 1989). In addition, their personal and social values have a strong influence on their daily assessment decisions via reflection (McMillan, 2003; Van Manen, 1977; Zeichner & Liston, 1987). Finally, *scripts* help reflective teachers to reflect by allowing them to

multi-task, for example, checking student understanding and lecturing at the same time, and guide teachers' thinking processes (Resnick & Klopfer, 1989).

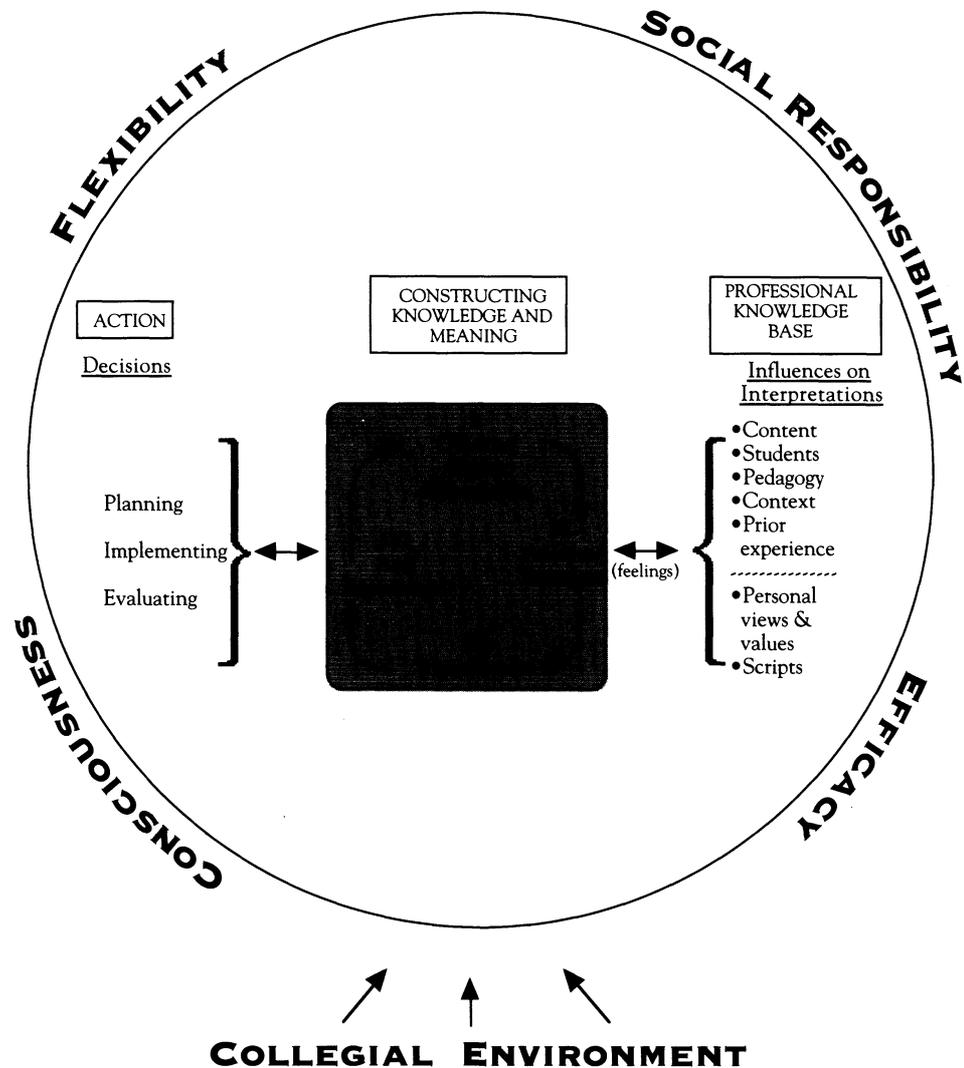


Figure 2.3. Framework for Teacher Reflection.

Source: Colton & Sparks-Langer, 1993.

Feelings, as portrayed in Figure 2.3 function as a bridge connecting professional knowledge and construction of that knowledge. Teachers' ability to reflect is deeply affected by their feelings (Harvey, Hunt, & Schroder, 1961). *Constructing professional knowledge* is a socio-cognitive process involving several stages. First, teachers encounter a situation and choose to completely or partially take part in it, then they

research the situation more clearly from many possible resources. Second, they analyse the research information to develop a theoretical basis to define the situation. Third, after the situation is clearly understood, teachers construct possible hypotheses that provide explanations of the situation and guidance for further actions to be taken. Once possible consequences of further actions are thoroughly considered, an action plan is carried out. Colton and Sparks-Langer (1993) also note, 'at every stage of the process, reflective decision-makers use automatic metacognitive scripts to guide their analyses and interpretations of situations, mental assessment of all possible hypotheses, selection of the action, and assessment of the final decision' (p. 49).

Given that the focus of teaching and assessment is all about learner outcome, and decision-making is often in association with *action* to take, Colton and Sparks-Langer propose three courses of action in the framework. As such, planning is conducted prior to teaching, followed by implementation of the teaching plan and then evaluation occurs after teaching is complete. In addition, the researchers note that teacher reflection should be encouraged in *collegial trusting environments*. It is also added that reflective teachers should drive their own attributes to engage more in the process of decision-making.

Overall, the framework proposed by Colton and Sparks-Langer (1993) provides a strong theoretical foundation for understanding flexibility in teacher assessment decision-making (i.e., what reflection is, how teachers reflect and the role of reflection in decision-making). Although this framework aims to umbrella a broader area of classroom assessment, the underlying theoretical values and coverage can be still applicable for this study to understand part of the teachers' decision-making process in the narrower area of oral language assessment.

Moderation: At key points in the process of teacher decision-making, moderation plays a vital part in assuring consistency and trustworthiness of the process. As defined by Maxwell (2010), ‘Moderation is a process for producing consistency across assessors in qualitative judgements of student performance or achievement’ (p. 457). Maxwell (2002) also notes that the term can be defined in two different dimensions, accountability and improvement. In the former, moderation functions as confirmation of informed assessments that are reported or publicised. In the latter, moderation helps develop assessment literacy for teachers so that they can make consistent and comparable assessment decisions. The choice of which dimensions to use depends on the purpose of assessment. With the notion of assessment for learning, moderation influences teacher assessment decision-making by helping teachers to either confirm or adjust their initial assessments. The role is well established in the literature on teachers’ assessment decision-making (Connolly, Klenowski, & Wyatt-Smith, 2012; Hipkins & Robertson, 2011; Klenowski & Adie, 2009; Maxwell, 2002), usually defined as a socio-cognitive process in which teachers are involved in sharing their initial judgements and working collaboratively with other teachers to reach a mutual understanding and perception of the quality of student output.

While moderation is useful in teacher decision-making, concerns regarding moderation have drawn a great deal of research attention. Hipkins (2010a) suggests two main perspectives on moderation. In the first, moderation is used to promote accountability in standards-based assessment decisions. This is the case in high-stakes assessment in which the results have individual, public and institutional consequences (Maxwell, 2010). The role of moderation in this regard is to evaluate judgements to ensure the trustworthiness of the entire assessment system. In the second perspective, moderation is suggested to be used as a source of professional learning activities (Hipkins,

2010a; Hipkins & Robertson, 2011). Moderation, in this regard, aims to ‘raise assessment quality and consistency in the longer term through professional development processes’ and therefore, is ‘more realistic, affordable and effective, especially where the emphasis is on formative assessment’ (Maxwell, 2010, p. 457).

Unlike flexibility, which is internal to individual teachers, moderation is more of a social activity. *Social moderation* involves teachers sharing and discussing their initial judgement on student output to establish a shared understanding (Gipps, 2012). In social moderation, teachers participate in different processes in which they may deconstruct, reconstruct or co-construct their perceptions, knowledge and skills (Hipkins & Robertson, 2011). Apart from its instant focus on improving consistency of teacher decisions, moderation can potentially help teachers to develop new knowledge and approaches of effective teaching to ensure student outcomes. Thus, moderation in teacher assessment decision-making has an effect on learning (Timperley, 2008).

Social moderation takes place in different forms at various points in the teachers’ decision-making process; however, Klenowski and Adie (2009) suggest three types of social moderation. The first is called the calibration model. In this model, teachers individually assess a student performance sample. Then, with the aim of reaching a consensus and shared understanding of student output, teachers discuss with each other their assessments. This moderation type occurs before teachers finish the assessment of the rest of the students in their classes. The second type of moderation is the conferencing model in which teachers assess some of the student samples and then select some of the assessed samples to share with others. After a shared understanding is reached, teachers may need to revisit all their assessments. In the last moderation type, the expert model, teachers are required to assess all performance samples and submit

some to an expert assessor to check if the standards are interpreted and applied consistently (Queensland Studies Authority, 2007, as cited in Klenowski & Adie, 2009).

When making moderation decisions, teachers may draw on a wide range of referents including concrete referents and social knowledge as referents (Hipkins & Robertson, 2011). The former consists of the standards themselves, guidelines and other professional advice and other aspects of student work. Specifically, standards are referred to as requirements and expectations about student performance, helping teachers to concentrate more on the assessment aim. Teacher guidelines and other professional advice can be used to assist standards referents by guiding the making of components of judgement or the entire judgement. The last concrete referent is other aspects of student work. When making moderation decisions, teachers may sometimes focus on individual components of student work, on an overall judgement associated with several components of student work or on weighing those different components.

Social knowledge—knowledge and beliefs—as referents comprise three significant categories including knowledge and beliefs about assessment, students and the intended curriculum. In teachers' knowledge and beliefs about assessment, the assessment approach can be decided by their prior experience of making judgements. In addition, their views about the consequences of assessment can inform their levels of decision. Teachers will not make a decision when they are aware it discourages students (Hipkins, 2010b). Every decision they make should be beneficial to students (McMillan & Nash, 2000). They are also sceptical if the assessment task meets any standards. Teachers' knowledge and beliefs as well as expectations seem to influence their moderation decision. However, there is a demand to draw a line between decisions based on student achievement and those based on student attitudes (Klenowski &

Wyatt-Smith, 2010). Finally, the effect of knowledge and beliefs about curriculum on moderation and decision-making is well documented.

In general, as a socio-cognitive process, moderation is reported to help improve and ensure the quality of teachers' assessment decisions. There are several types of moderation; therefore, depending on the purpose and nature of assessment, teachers can decide which to use. During moderation, teachers may consider several different referents to make decisions, hence knowing these will help teachers to reach a collective understanding of student work and make consistency in their decision possible. Moderation can take place before, during and after the decision-making process and it may be useful when moderation is carried out on an ongoing basis (Maxwell, 2002).

2.7. A Need for a Conceptual Framework for Teacher Language

Assessment Decision-Making

In this section, research gaps are identified and embedded in the call for a conceptual framework to better understand teacher decisions in classroom language assessment.

Given that whenever and wherever humans, whether they be teachers or raters, are included in the language assessment process, the nature of such assessment is subject to inherent variability (Davison & Leung, 2009). For decades, variability in language assessment has been an issue of concern and a stimulus for a wide variety of research. Although this research area has been extensively and thoroughly investigated and rater training has been suggested as an effective solution for minimising rater variability, the majority of such research has been done in settings of large-scale testing like IELTS (Carey et al., 2011), TOEFL (Xi, 2010; Xi & Mollaun, 2009, 2011), MELAB and so forth. However, most of the schooling process is constituted by day-to-

day assessments by classroom teachers of their own students. As a global tendency, this kind of assessment is gradually being adopted in formal assessment (Cumming & Maxwell, 2004; Davison, 2004; Government, 2005; Hamp-Lyons, 2009; Pryor & Lubisi, 2002; Xu & Liu, 2009). Teachers are assumed to be the best people to assess their students (William, 2001), yet this assumption may also be perceived a source of threats to the trustworthiness of such assessment because, as stated earlier, subjectivity is an inherent characteristic of human assessors. The issue of subjectivity in teacher-based assessment is even more of a concern when experienced teachers assess unfamiliar students using assessment tools they do not know.

With the attempt to deal with or alleviate variability in teacher-based assessment, much research has been done to conceptualise teacher assessment practices to better understand what they do in the black box—the classroom. Since the early 1990s, a range of conceptual frameworks has been developed and introduced. For example, Colton and Sparks-Langer (1993) proposed a conceptual framework guiding the development of teacher reflection and decision-making. Smith (1996) conceptualised a framework to support teacher ESL decision-making. In another example, McMillan (2003) also conceptualised the teacher assessment decision-making process and proposed a range of implications for classroom assessment theory and practice. In addition, most recently, to promote effective teacher training and instruction in the area of science, Clough, Berg, and Olson (2009) also proposed a framework for understanding teacher decision-making.

Overall, it seems that teacher assessment decision-making has been well supported. However, these conceptual frameworks are developed to support general teacher assessment decision-making practices, rather than focusing on a specific dimension (e.g., language assessment of such practices).

Furthermore, another research gap involves the depth of research into the variability of teacher assessment decision-making. Almost all previous studies investigating the issues of variability in teacher language assessment have dealt simply with the phenomena of teacher variations in their judgements. Notably, scores are used as the primary source of data analysis. When only scores are used to compare rating practice or to determine variability in their assessment, it is predictable that differences will be found. It is necessary to further examine the nature of these differences. In addition, day-to-day classroom assessment does not always necessarily deal with scores. Sometimes the purpose of an assessment task or activity is to know whether students have reached the goal of a learning unit or to identify students' strengths and weaknesses so that appropriate adaptations can be done to support learning. Yet, research has not investigated the variability of the process of how teachers perceive and interpret their students' performance responding to an assessment task, and how they then make judgements based on their perceptions and interpretation. Therefore, there must be a call for research to uncover the process of teacher decision-making.

These two research gaps suggest an urgent need for a conceptual framework that is closely aligned to teacher language assessment decision-making practices and that sheds more light on this socio-cognitive process. This proposed framework would make variability in teacher assessment decisions no longer a critical concern, because once the socio-cognitive process of teacher assessment decision-making is more thoroughly understood, a better course of action would be taken to tackle concerns over the quality of teacher-mediated language assessment that has long been known for variability and inconsistency.

2.8. Summary

This chapter has reviewed the literature related to the variability of teacher oral assessments and the influential factors that are relevant to the aims of this study. In the Australian as well as the global context, there are increasing calls for the standardisation of ESL/EAL/D assessment. However, when the number of ESL/EAL/D teachers is increasing because of a dramatic growth in the number of students, a common concern among language educators and stakeholders is the quality of those teachers. This, in conjunction with the lack of standardised assessment sources, may result in variability in teacher assessment practices. Since teachers are the centre of assessment and the main source of variability, understanding the nature of their variability is important when exploring their practice and influential factors. As assessing oral performance is a critically complicated socio-cognitive process, understanding what teachers understand and do when making assessments and their awareness of the factors affecting their understanding and decisions may help to improve their assessment practices and build a more trustworthy assessment system. In this study, a mixed method research approach is adopted to investigate the process teachers make in their assessments of students' oral performance in an Australian context. The details on the research methodology are presented in the next chapter.

Chapter 3. Methodology

3.1. Introduction

In Chapter 2, the key areas of the investigation related to this study were identified. Specifically, the theoretical frameworks that shapes classroom assessment practice were critically reviewed and discussed. Notably, variability in language assessment—the key focus of this research study—was also discussed at length along with teacher differences in making assessments of students’ oral language communication skills and influential factors in association with teachers’ background and other assessment factors. This study is designed to investigate how teachers make decisions when they assess students’ ability to communicate orally and what affects their decision-making practice, including how they justify the decisions they have made. Such qualitative focuses will yield rich information beyond the information gathered from their assessment scores.

In this chapter, the aims of this study are outlined and operationalised into the research questions. Next, a mixed methods research approach is presented together with its characteristics, strengths and weaknesses, and the rationale for the use of the mixed methods approach to answer these research questions. The methodology section is followed by the context of the study, including a description of the development of the teacher-based oral assessment system that provided the resources for this study and the EAL/D instruction and assessment practice in New South Wales (NSW) where the study was conducted. The research design is also presented at length, detailing the design and implementation of the tools for data collection. There are four data collection methods in this study: (1) a questionnaire, (2) a teacher-based assessment activity, (3) a

retrospective think-aloud protocol and (4) follow-up interviews. In addition, in this chapter, information about how the methods of data collection were piloted is provided, followed by detailed descriptions and explanations of how the quantitative and qualitative data were analysed.

As stated in Chapter 1, this study is part of a larger research project developed to enhance assessment literacy for EAL/D teachers contributing to building a trustworthy language assessment system for classroom settings. This particular study was undertaken to investigate how teachers make assessments of their students' oral language work and to explore factors which influence their assessments. More importantly, this study investigates their process of assessment decision-making. The research questions are:

1. To what extent are teachers' assessments of students' oral English communication skills consistent with one another?
2. What are the factors that influence teachers' assessments?
 - a. What factors related to teachers' background influence their assessments?
 - b. What factors related to the assessment tasks affect teacher's assessments?
3. What are the characteristics of teacher assessment decision-making?

It is clear that the data needed to answer the first and the second research questions are very quantitative in nature. For this reason, a questionnaire was designed to collect demographic data from teachers, along with the use of documents about the tasks and students to explore how teachers' assessments are influenced by the key factors identified in the literature review. A teacher-based assessment activity was also designed to examine the consistency of teacher assessments. Data obtained from the questionnaire and the assessment activity along with the use of documents help to

explore the factors influencing teacher assessments. Answering the second question also requires qualitative data about teachers' justifications for their assessment decisions. Such data are collected via the retrospective think-aloud protocol. The last research question is aimed at identifying characteristics of teacher decision-making, again mainly qualitative data from the think-aloud protocol.

As Dörnyei and Taguchi (2010), Dörnyei (2007), Johnson and Christensen (2010), Johnson and Onwuegbuzie (2004) and Teddlie and Tashakkori (2011) note, research questions decide the choice for research approaches. To best answer these research questions, it is important to adopt an approach to research that addresses both quantitative and qualitative components.

3.2. Research Approach

In this section, a brief review of mixed methods research is presented, to explain what mixed methods approaches are (including strengths and weaknesses) and how and why these are important in doing research.

3.2.1. Pragmatism: a philosophical partner to mixed methods research

During the 1870s, a philosophical movement claimed that an ideology or proposition is true if it works satisfactorily. This movement is called philosophical and methodological pragmatism. By taking a pragmatic method or maxim to discover the meaning of, or to make judgements about, an idea or a phenomenon, people must take into consideration its practical consequences. To better understand a real-world phenomenon, people take purist philosophical positions, whereas, it is argued that 'if two ontological positions about the mind/body problem (e.g., monism versus dualism), for example, do not make a difference in how we conduct our research then the

distinction is not, for practical purposes, very meaningful' (Johnson & Onwuegbuzie, 2004, p. 17). Quantitative and qualitative research approaches are characterised by a set of beliefs and each approach has its own pros and cons. Therefore, it is appropriate to conduct quantitative research in some situations and qualitative research in others. The mixed methods approach from a pragmatic position fits 'together the insights provided by qualitative and quantitative research into a workable solution' (Johnson & Onwuegbuzie, 2004) and 'helps answer the questions that we value and provides workable improvements in our world' (Onwuegbuzie & Johnson, 2006, p. 54). Mixing should be carried out in a way in which the best opportunities are given to the best answer research questions.

3.2.2. Mixed methods approach

Mixed method research is viewed as a 'new third chair, with qualitative research sitting on the left side and quantitative research sitting on the right side' (Johnson & Onwuegbuzie, 2004, p. 15). Debates over research methods and approaches mean mixed methods research is also seen differently by leaders in the field. For example, it is called ethnographic residual analysis by Fry, Chantavanich, and Chantavanich (1981), blended research by Thomas (2003), integrative research by Johnson and Onwuegbuzie (2004), multi-method research by Hunter & Brewer (2003) and Morse (2003) and mixed research by Johnson and Christensen (2010). Since mixed methods research is defined in several ways, Johnson, Onwuegbuzie, and Turner (2007) refine definitions of the research methods under five themes: 1) what is mixed, 2) the mixing stage, 3) breadth, 4) why and 5) orientation of mixed research. In terms of what is mixed, it is widely agreed that mixed methods research combines both quantitative and qualitative research. Regarding the mixing stage, debates remain over when and where it is mixed.

Some suggest that, by definition, mixing takes place when researchers collect data. Conversely, others claim that mixing occurs not only in the data collection stage but also in the data analysis stage. However, it is more commonly agreed that research is mixed at all stages. Breadth is considered an underlying theme of the first two themes. A very significant theme related to purposes is why people need to be mixing when conducting research. Mixed methods research is commonly employed because it enhances description and understanding and, more importantly, helps with the triangulation of research findings. The last theme is the orientation of mixed methods research and describes how the research approach is driven. One orientation is called ‘bottom-up’, meaning the mixed methods research approach is driven by the research questions (Johnson & Onwuegbuzie, 2004; Johnson et al., 2007). Another orientation is known as ‘top-down’ and suggests that the research questions do not determine the mixed methods approach but the researcher’s quest does (Johnson et al., 2007). Taking all these themes into consideration, Johnson et al. (2007) propose a composite definition of mixed methods research:

Mixed methods research is the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration (Johnson et al., 2007, p. 123).

Johnson et al. (2007) also classify mixed methods into three different subtypes (see Figure 3.1). The first subtype is called *equal status*. Accordingly, both qualitative and quantitative data and approaches are purely mixed and believed to equally add to the insights in workable solutions. In another subtype of mixed methods research called *qualitative dominant*, a researcher relies mainly on qualitative, constructivist-

poststructuralist-critical perspectives and uses quantitative data and approaches for additional benefits. Conversely, the *quantitative dominant* mixed methods research subtype suggests a scenario in which a quantitative researcher may also realise the benefits of using qualitative data and approaches.

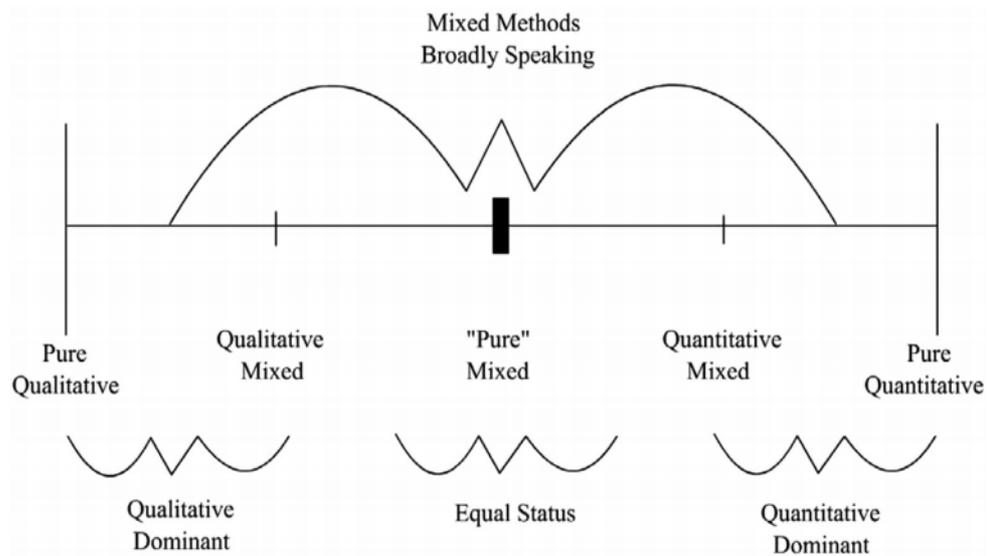


Figure 3.1. Subtypes of Mixed Methods Research.

Source: Johnson et al., 2007, p. 124.

Like other research approaches, the mixed methods approach has several characteristics. The first characteristic is called *methodological eclecticism*. By rejecting the idea of incompatibility that qualitative and quantitative research should not be mixed because they are epistemologically different in approaches, Teddlie and Tashakkori (2011) suggest that the integration of qualitative and quantitative methods is good because it cancels out or minimises respective weaknesses (Johnson & Onwuegbuzie, 2004). This characteristic also implies that researchers are free to mix research in a way their research questions are best answered, for example, the use of diverse data sources (Schulenberg, 2007). The second characteristic of the mixed methods approach is *paradigm pluralism*, literally understood as adopting a multiple research approach to

philosophically support the use of the mixed methods approach (Teddlie & Tashakkori, 2011). Another characteristic of mixed methods research is ‘a celebration of diversity at all levels of the research enterprise from the broader more conceptual dimensions to the narrower more imperial ones’ (Teddlie & Tashakkori, 2011, p. 287). Hence, the diversity of mixed methods research is demonstrated in that this approach addresses both confirmatory and exploratory questions. Moreover, as mixed methods research differs from a single method research approach in that it is *an initiative and cyclical approach to research*. In one study, a researcher can include both deductive and inductive logic (Krathwohl, 2004). Specifically, a researcher can make generalisations from facts or observations (i.e., inductive logic) and from these generalisations can then make tentative hypotheses or predictions of events (i.e., deductive). Placing *an emphasis on research questions to determine research methods* is the fifth characteristic of the mixed methods approach (Bryman, 2008; Johnson & Onwuegbuzie, 2004; Teddlie & Tashakkori, 2011). The centrality of research questions is that researchers starts conducting a research project by identifying a research problem that may be conceptual or practical, and then identifying the purposes of doing research. These purposes are operationalised into research questions that determine which instruments are to be used (Johnson & Christensen, 2010; Johnson et al., 2007; Teddlie & Tashakkori, 2011). Mixed methods research is also characterised by *a set of basic research designs* such as parallel missed designs and triangulation designs. In the former, mixing takes place independently ‘either simultaneously or with some time lapse’ (Teddlie & Tashakkori, 2009, p. 341) and in the latter, mixing helps seeking ‘convergence and corroboration of results’ (Johnson & Onwuegbuzie, 2004, p. 22).

By taking the characteristics of mixed methods research into consideration, this study aims to understand the classroom assessment phenomenon in which teachers

make assessments against students' oral language communication performance.

Teachers' assessment may be different, driven by a wide range of factors related to the teacher themselves, students and assessment tasks. To reach a complete understanding of teacher assessment variability, it is necessary to gain insight into the process teachers use to make decisions about students' language development by examining what they do when they assess and by identifying what shapes their judgements.

Like any single method research, mixed methods research has several strengths and weaknesses. Weaknesses in qualitative research may be compensated by strengths in quantitative research; therefore, in sitting on a new third chair, mixed methods research entails all the strengths of qualitative and quantitative research. Moreover, the mixed methods approach to research is that non-numerical data such as words, images or narrative are employed as meaningful additions to the numerical data, making numbers more informative. In return, numbers help add accuracy to non-numerical data (Johnson & Onwuegbuzie, 2004). Another advantage of mixed methods research is inherent to its characteristic as an initiative and cyclical approach to research. In particular, using this approach to a study, researchers can inductively make a generalisation and deductively examine a grounded theory (Johnson et al., 2007; Teddlie & Tashakkori, 2011). Further, researchers adopting a mixed methods approach can answer different types of questions from the more general and conceptual dimensions to the narrower and more imperial dimensions (Teddlie & Tashakkori, 2011), because they are relatively flexible in choosing their research methods. Furthermore, the combination of qualitative and quantitative approaches to research can strengthen evidence to draw a conclusion and; therefore, increase the generalisability of results (Johnson & Onwuegbuzie, 2004).

However, although in the mixed methods approach, as stated earlier, the strengths of quantitative research approaches can overcome the weaknesses of qualitative research approaches and vice versa, the mixed methods approach entails its own weaknesses. Firstly, because it is a combination of qualitative and quantitative research approaches, conducting such a mix may cause certain difficulties for single researchers. It is particularly difficult when they must implement both qualitative and quantitative components at the same time. Further, the mixed nature of this approach to research suggests that researchers are required to have a complete understanding of different methods and approaches so that they can mix with confidence. For this reason, at some stage, researchers are recommended to ‘always work within either a qualitative or a quantitative approach’ (Johnson & Onwuegbuzie, 2004, p. 21). Another weakness of mixed methods research is in relation to cost and time. A mixed methods study is always more expensive and time-consuming than a single method study. Finally, debates over critical issues related to this third new chair remain unsolved. Research methodologists should work collaboratively to resolve paradigm mixing problems, qualitative analysis of quantitative data and interpretation of conflicting results.

3.2.3. Justification for mixed methods research

In this study, the mixed methods approach to research is employed for the following reasons. First, this research approach can address ‘different research questions’ (Bryman, 2006). As suggested by a range of research methodologists, choosing a research approach in a study depends on the research questions that the study aims to answer (Dörnyei, 2007; Dörnyei & Taguchi, 2010; Johnson & Christensen, 2010; Johnson & Onwuegbuzie, 2004; Johnson et al., 2007; Teddlie & Tashakkori, 2009, 2011). This study aims to examine oral language assessments made by classroom

teachers and explore the factors that shape their assessments and the characteristics of their assessment practice behaviour. The approach to the first research question aimed at examining teachers' oral language assessment is designed to collect data in numerical form. Specifically, teachers are asked to make judgements of students' speaking samples and these judgements are scored in numbers using an assessment rubric. This is considered quantitative because, as Aliaga and Gunderson (2000) note, such research seeks to explain a phenomenon by the use of numerical data that are mathematically analysed. In contrast, the second research question regarding the characteristics of assessment practice behaviour of teachers is more qualitative. Accordingly, this question aims to further examine the process teachers use to make their judgements and; therefore, their thoughts and perceptions while they assess are important. To collect this type of information, the think-aloud method is adopted. Finally, addressing the last research question: 'What are the factors that influence teachers' assessments?' requires both quantitative and qualitative data that are designed to be collected through a questionnaire and documents. While the purpose of the questionnaire in this study is collecting information from teachers on their demographic and academic backgrounds, the documents are expected to provide details about students and tasks. Both the data collected from these tools are in numerical and non-numerical forms.

Another reason for adopting the mixed methods approach in this study is the inseparability of 'completeness' and 'explanation' (Bryman, 2006). Completeness 'refers to the notion that the researcher can bring together a more comprehensive account of the area of enquiry in which he or she is interested if both quantitative and qualitative research are employed' (p. 106). In this study, the quantitative component is designed to describe the phenomenon of how teachers assess the speaking language performance of students. However, the description is formed based on observations of

teachers' actual assessment performance and, thus, fails to present a more comprehensive description of the phenomenon. Therefore, the qualitative research component of this study is critical to address the research questions fully. At the same time, the qualitative research component also seeks to explain the phenomenon identified by the quantitative component. After teachers' assessments are examined and patterns of assessments are found, it is necessary to further explore why those patterns are formed.

For all the above-stated reasons, a mixed methods research approach was adopted in conducting this study. Specifically, a *sequential* and *equal status* mixed methods design was employed, with both qualitative and quantitative components equally weighted, and the two research components sequentially conducted—the quantitative component conducted first, followed by the qualitative one.

3.3. Context of the Study

Although this study was conducted in NSW, its research tools and materials, as introduced in Chapter 1, were adapted from a larger project in Victoria. For this reason, to provide a better and clearer understanding of this study's context, a brief description of the Victorian project will first be presented, starting with the purpose of the project, its components and the process of developing the prototype teacher-based oral assessment system which provided the assessment materials for this study, followed by a full description of the context in which all data were collected.

The aims of the TEAL project included enhancing the assessment literacy and competence of teachers by building an assessment toolkit developed by and for teachers who are working with EAL/D school-age learners. The TEAL project consisted of three main components: (1) the development of a web-based ESL assessment resource centre,

(2) development of a prototype teacher-based writing and oral assessment system and (3) development of computer adaptive tests of reading and vocabulary. These were briefly mentioned in Chapter 1. Since this study aims to make investigate teachers' oral assessment decision-making, only the TEAL system of teacher-based oral assessment is presented at length.

3.3.1. Development of the Teacher-Based Oral Assessment System

The aim of the TEAL teacher-based assessment system is to provide an assessment training milieu and a professional forum to help teachers practice assessing with confidence before and even while assessing their students in classroom contexts. Exemplar oral assessment systems are designed to provide diagnostic information on students' English language and literacy development, as well as information on their level of accomplishment aligned with the various stages of Victorian EALD curriculum documents. To help with the development of the teacher-based oral assessment system, a bank of common oral tasks has been developed with the input from teachers and is made available on the project website, see <http://teal.global2.vic.edu.au/>, for participating teachers to view and provide feedback. The set of 21 tasks are classified into three different genres including *informative* (tasks 1, 2, 6, 9, 10, 11, 14, 16, 19) *imaginative* (tasks 3, 8, 12, 13, 17) and *persuasive* (tasks 4, 5, 7, 15, 18, 20, 21). They are also classified into three types of oral communication: (1) listening and responding, (2) interaction and negotiation and (3) oral presentation. The project uses the following criteria to evaluate oral assessment tasks: (1) being at the right level for students, (2) being built on previous learning, (3) being intrinsically motivating to maximise student participation, (4) being authentic in purpose, (5) reflecting the learning outcomes at the

appropriate stage of the unit of work, (6) being familiar to the students and (7) being a stepping stone to the next learning goal.

The development of the tasks was followed by the generation of videos of sample students' oral work designed for teachers to practice and improve their assessment knowledge and skills. A range of students were asked to respond to the tasks individually, in pairs or in groups. Student responses were filmed, and the output edited and published on the project website, along with the tasks. At the same time, to facilitate teachers' assessment practice, a set of assessment criteria were also developed by TEAL experts and trialled and evaluated by participating teachers. The task assessment criteria use a four-point scale rating rubric that includes equally weighted categories. Teachers provide their feedback on the criteria through workshops by viewing sample videos of students working on one of the tasks. In addition, participating teachers are provided with full access to the videos of speaking samples and criteria so that they can practice assessing at their convenience. While assessing the videos, teachers record their judgements by filling out an assessment sheet designed by TEAL. Teachers are also advised to align students' levels of task completion with the AusVELS EAL/D continuum. Teachers are encouraged to discuss their assessments with each other via workshops provided by the project or via a password-protected discussion forum of the project to agree upon common assessments. These common assessments are used as benchmark assessments and teachers are encouraged to refer to these when they assess their students.

3.3.2. EAL/D learning and teaching in NSW

In NSW, according to the Department of Education, it is compulsory for schools to provide a range of support to EALD students, especially to newly arrived students so

that they can succeed at school and reach their full potential. Schools must ensure that EAL/D support programs are included as an integral part of their plan and that EAL/D teaching positions are filled by teachers with relevant EAL/D teacher qualifications. Furthermore, procedures for the identification, assessment, reporting and tracking on EAL/D students must be in place and maintained (NSW Education, 2017a). To reinforce implementation of EAL/D support, the Department has issued additional policies and guidelines to help the schools, for example, the Multicultural Education Policy, English as an Additional Language or Dialect: Advice for Schools, EAL/D School Evaluation Framework (NSW Education, 2017b), Checklist for Effective EAL/D Student Support, EAL/D Education (NSW Education, 2014) and the like. In addition, the Multicultural Education Policy provides guidelines to respond to the diversity of culture, language and religion in NSW schools. Schools must provide opportunities to help students to achieve their educational and social outcomes and take part confidently in the cultural diversity of society. In addition, English as an Additional Language or dialect: Advice for Schools (NSW Education, 2014) provides a set of advice and guidelines to enable schools to deploy appropriate EAL/D student support. Additionally, teachers are also advised to use the ESL Scale (NSW Education, 2006) in conjunction with the EAL/D syllabus to address EAL/D student needs and to help them have access to English curriculum outcomes and content. The ESL Scale, according to the NSW Education Standards Authority, provides a description of English language learning progression for EAL/D learners.

Under the Australian national education agenda, NAPLAN (National Assessment Program, 2017) standardised tests are used to assess student competence in English literacy and numeracy at Years 3, 5, 7 and 9. However, newly arrived EAL/D students are not assessed by NAPLAN and, as mentioned above, there is a shortage of

resources for teacher support in EAL/D assessment (Davison & Michell, 2014). This study aimed to provide some insights into how assessment support material developed for one educational system could be used in another.

3.4. Research Design

This study was conducted in two stages. In the first stage, quantitative data was collected to examine assessments made by teachers and to explore how their assessments were influenced by the factors relating to teacher background, student background and assessment tasks. Details of the research design are presented in Figure 3.1.

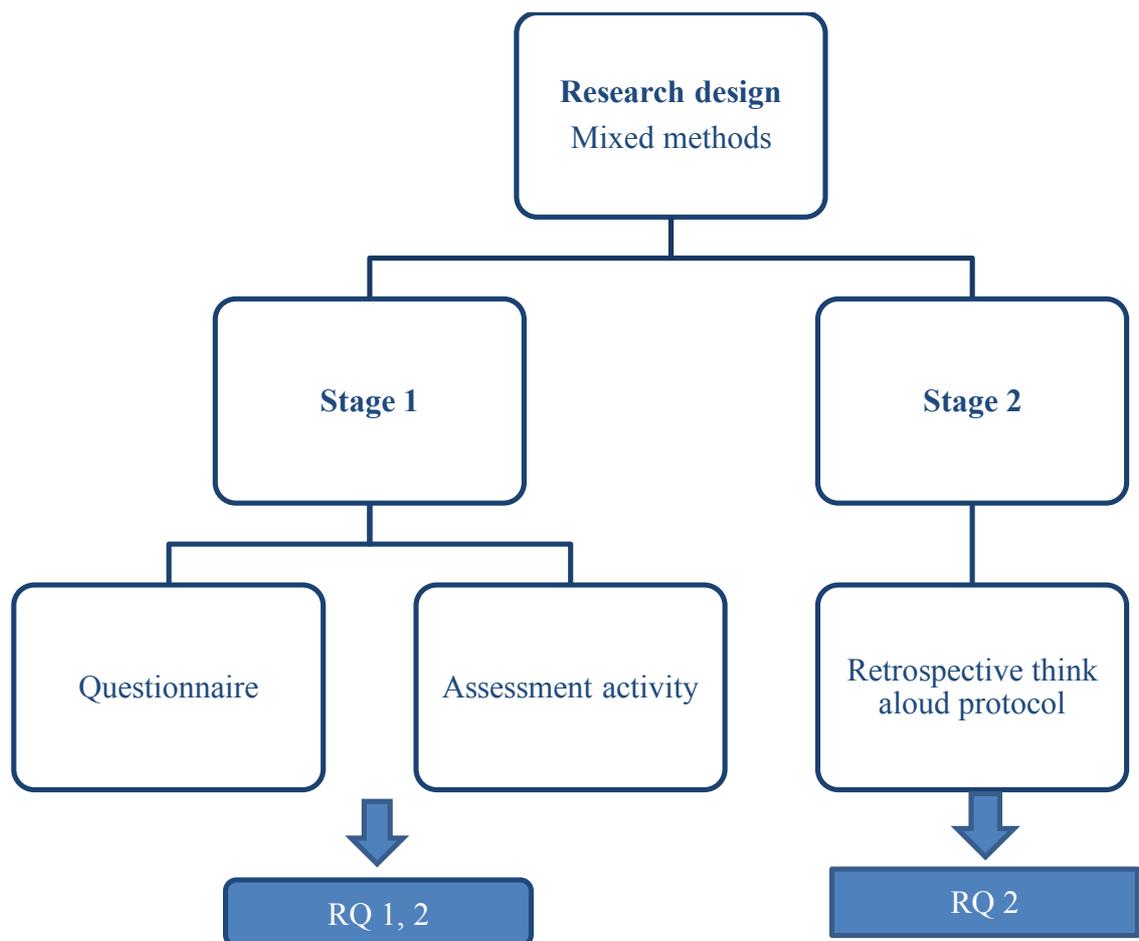


Figure 3.1. Research Design.

In the first stage, a questionnaire was used to collect demographic information from the participating teachers, to be later correlated with the scores they assigned to specific samples of student oral language use taken from the TEAL website. After answering the questionnaire, the participants were asked to take part in a teacher-based assessment activity in which they watched a set of videos of students' performances and assigned scores to students' performances. Student background information (e.g., gender and first language) and information about the task (e.g., types and genres) were also collected from task annotations and an analysis of videos that are publicly available on the TEAL project website.

The second stage of the study was designed to collect qualitative data to further explore how teachers' assessment decisions were shaped by a range of factors. The retrospective think-aloud activity was used to obtain information about how the teacher assessments were produced. To examine further explanations of teachers' decisions and justifications, follow-up semi-structured interviews were conducted. Details of the research tools and the data collection will be provided in following sections in this chapter.

3.5. Stage 1

3.5.1. Participants

This study focused on examining teacher assessments of oral English communication skills and exploring the factors which influenced those assessments. Because this study was set in the context of EAL/D learning and teaching in NSW, it involved EAL/D teachers taught in that area. Teachers were drawn from various backgrounds and different schools and locations. The only criterion for participant selection was that the participants had to be EAL/D teachers or ESL teachers who were

currently teaching EAL/D. The targeted participants were all teachers who were teaching ESL and EAL/D at primary and secondary levels in all three sectors: public schools, independent schools and Catholic schools in NSW.

Participants were selected using convenience sampling. An advertisement was posted via the website of the NSW Association of Teachers of English to Speakers of Other Languages (ATESOL). Participants for the first stage were recruited to provide two sources of data. However, those participants who provided information in the first stage could withdraw or refuse to give further data in the second stage if they wished. All participants were invited to the next stage, that is, the interviews. Again, they were free to decide whether they wanted to proceed or stop.

Twelve teachers took part in the full research study. As mentioned previously, one of the aims of this study was to investigate whether teachers' professional and personal background information influenced their assessment decisions. The demographic information about the participants collected from the questionnaire is shown in Table 3.1.

Table 3.1

Teachers' Demographic Information

	Age	Current teaching position	Specific TESOL qualifications	Experience (years)	Main language exposure
T01	56+	Others	Yes	16+	Chinese
T02	41–55	Secondary	Yes	11–15	Chinese, Korean, Vietnamese
T03	26–40	Secondary	Yes	16+	Chinese
T04	26–40	Primary	Yes	11–15	Thai, Chinese, Arabic

T05	41–55	Secondary	Yes	16+	Vietnamese, Arabic
T06	26–40	Others	Yes	6–10	English
T07	56+	Others	Yes	16+	English
T08	56+	Others	Yes	16+	English
T09	56+	Primary	Yes	16+	Chinese, Arabic, Persian
T10	41–55	Primary	Yes	11–15	Chinese, Spanish
T11	41–55	Secondary	In progress	11–15	A range of languages other than English
T12	56+	Primary	Yes	11–15	Hindi

Note. T: Teacher

As can be seen from Table 3.1, all the participants were female ($n = 12$). Three out of the 12 were aged between 26–40 years old, four were aged between 41–55 years old and five were aged 56 years and above. In terms of languages spoken, while the majority (11 out of 12) indicated that they spoke English at home, one said her home spoken language was Mandarin. In addition, 75 per cent of the participants indicated that that they spoke English outside their home on a regular basis, while the rest communicated in Filipino, Tagalog, French, Indonesian, Japanese and German.

Regarding their teaching profile, teachers were equally split into three groups. Specifically, four out of 12 were teaching EAL/D at primary level, four at a secondary level and the other four were not currently teaching EAL/D at the time of responding to the questionnaire. Teachers in the last group were EAL/D specialists; however, they were involved in management and consultation in EAL/D-related teaching and learning activities. However, they worked closely with EAL/D learners at both the primary and secondary levels. 58 per cent of the teachers were working in the government school

sector, 17 per cent in the Catholic school sector and 8 per cent in the independent school sector.

All the teachers were qualified in teaching EAL/D in addition to their general teaching qualification; only one was still studying towards their TESOL qualification. In terms of teaching experience, they were very experienced EAL/D teachers. Half had been teaching for more than 16 years, three had been teaching for between 11 to 15 years and three had been teaching for between five to 10 years.

When asked about the main language groups they worked with, the teachers indicated that they were working with students from quite diverse language backgrounds (e.g., Chinese, Vietnamese, Arabic, Persian, Korean, Hindi, Thai, Khmer, and Spanish).

3.5.2. Questionnaire

Questionnaires are one of the most commonly used methods to collect quantitative data. Questionnaires are used to measure various types of individual characteristics such as thoughts, beliefs, attitudes, values, perceptions and behavioural intentions (Borg, 2012; Marshall & Rossman, 2010). A questionnaire can also be used to collect factual data such as personal background and demographic information, attitudinal data such as opinions, values and beliefs and behavioural data such as experiences and actions (Dörnyei & Taguchi, 2010).

Employing a questionnaire in this study was appropriate as I intended to collect data on demographic information that could be directly drawn from the participants. Besides, a questionnaire has ‘unprecedented efficiency’ in terms of cost and time. By administering a questionnaire, a huge amount of data can be obtained from a large group of participants in a short period of time (Dörnyei, 2007; Dörnyei & Taguchi,

2010). Data collected via questionnaire are also ‘versatile’ as it can be used ‘with a variety of people, in a variety of situations, targeting a variety of topics’ (Dörnyei & Taguchi, 2010, p. 6). Further, inclusion of open-ended questions in a questionnaire can be helpful in obtaining valuable information about people’s inner perspectives, especially when a variable is loosely defined (Johnson & Christensen, 2000), and closed-ended questions offer practicality and convenience in responding and analysis. Finally, questionnaires are considered a data collection approach that promotes security of anonymity (Dörnyei, 2007) and, thus, increases respondents’ honesty.

The questionnaire in this study was designed to simply collect data on teachers’ background, teaching positions and assessment experiences (see Appendix B). The questionnaire was constructed to collect information on key factors that the literature review suggested might have some effect on teacher assessment decision-making. Mostly close-ended questions were utilised in the questionnaire. Close-ended questions help to collect data on participants’ backgrounds in a less time-consuming way. After being developed, the questionnaire was pre-tested with a different group of five participants who shared several characteristics with the pool of research participants.

3.5.3. Teacher-based assessment activity

A teacher-based assessment activity was conducted immediately after the questionnaire administration. The activity was designed to examine the extent to which teachers’ assessments were consistent with each other. Although this teacher-based assessment activity was set up by a researcher (i.e., myself), it resembles what EAL/D teachers already do in their classrooms.

In this activity, participants were asked to assess samples of videos of student performance in conjunction with the use of other assessment materials such as assessment rubrics.

3.5.3.1. Tasks

One of the aims of this study was to identify factors at the teacher-, student- and -task-levels that are potentially related to teacher assessment outcomes. To serve this aim, my criteria in select relevant assessment tasks was to include several types of tasks such as listening and responding, interaction and negotiation and oral presentation and several genres such as informative, imaginative and persuasive. As indicated earlier, the three tasks were selected from the pool of 21 oral assessment tasks originally developed for the TEAL assessment project in Victoria. These tasks were originally designed to assess upper primary and secondary students' English language performances, meaning that both primary teachers and secondary teachers can suitably use these tasks to evaluate their student outputs. Detailed descriptions of the tasks are available in Appendix C but are summarised below.

Task 13: Choosing a gift for a character—This task assesses the ability of students to participate in a collaborative discussion with their peers, in which they discuss a character and events in a literary work they are familiar with, to reach agreement about a suitable gift for a character in the story. It provides assessment information about EAL/D students' abilities to negotiate with each other and discuss a literary work they have been studying.

Students watched the same film, *What's Eating Gilbert Grape?* (Hallström, 1993) prior to being given this task. The students had previously discussed the characters in the film and the idea of giving a gift to a character that would be useful or

suitable for them given the character's personality or circumstances in the film. In starting the discussion, the students were asked to say a little about the character they had chosen and what happened to them in the film, before discussing suitable gifts. The students were asked to discuss several possible gifts and to give the reasons why the gifts might be suitable, or not be suitable, before coming to a decision.

The video based on this task shows the discussion of a group of three students. One student is a female from China, another is a male with Somali as his first language and the third is a male whose first language is Arabic. All the students were in Year 10.

Task 19: A book or film review—This task is an informative task of the oral presentation type. The task assesses students' ability to describe the plot, characters, relevant themes and issues and provide evaluative comments and a personal response to the work. To conduct the task, students read a novel or view a movie. They are then asked to give a brief spoken report and evaluation of the work, in response to questions from a classmate or teacher.

The language demands of such a review can be complex and varied. A range of present and past tenses can be used in describing the plot, certain events in the work, the characters, themes and issues arising from the text and in giving a personal response. Some meanings require present tense, particularly the discussion of themes and issues. Recounting the plot and retelling events in the story can be achieved by using either the 'historic present', such as 'Paikea rides the whale', or past tense, such as 'Paikea rode the whale', when the plot is presented as a narrative. While either present or past can be used, there is an expectation of consistency in the use of one main tense once the retelling has begun and that the speaker will continue in the same tense. Similarly, characters can be described either in present or past tenses. Present or past tense can be used in giving a personal response to the work, for example, 'It's alright', or 'I thought

it was good'. The challenge for EAL/D learners is to use this range of tenses consistently in acceptable ways when giving the review.

The video generated from this task shows the conversation of two students, one male and one female. Both students speak Chinese as their first language. The students were in Year 8.

Task 21: Job interview role play—This task involves an interactive and relatively spontaneous performance in which students are interviewed about themselves in relation to a hypothetical job. It assesses several areas of English language use, including the use of simple present tense to talk about themselves, their qualities and attributes, the use of past tense or present perfect to talk relevant experiences and modal verbs to talk about the skills they have. It also assesses students' abilities to discuss hypothetical events.

The situation also requires students to use culturally appropriate ways of talking about themselves in a positive way, without being judged to be overconfident, conceited or bragging. Indeed, it is a delicate balance, for interviewees are expected to sound positive about themselves, yet not overly confident of their own abilities. The task also provides teachers with information about their students' fluency and spontaneity in an interview situation, in which they may be 'put on the spot' by unexpected or tough questions within predictable parameters.

The video generated from Task 21 is the performance of a male student in Year 8. He comes from Mongolia. The student is in role of the interviewee answering the male interviewer. The tasks were also contextualised with teachers' annotations on the tasks, videos of other student speaking samples and assessment templates. The task annotations provide information on tasks such as the types and genres that was then used for correlational analysis along with teachers' judgements. Videos and assessment rubrics and assessment templates were treated as supplements to the assessment activity.

3.5.3.2. *Assessment criteria*

Teachers were asked to score students' oral language development by using the provided assessment criteria that were adapted from the TEAL project. These assessment criteria were developed by a range of experienced teachers, and were piloted and validated with both primary and secondary school teachers. As this study was designed to evaluate the trustworthiness of teachers' assessments to see what kinds of support and adjustment were needed to improve teacher language assessment literacy and competence for the TEAL project, this study aimed to examine how experienced teachers who were used to working with similar kinds of assessment criteria used the TEAL rubrics for the first time. Each assessment task had a different criteria sheet indicating four levels of achievement across the EAL/D developmental continuum. The criteria contain detailed descriptors for each performance level, organised in four levels for four different linguistic categories such as communication, cultural conventions, linguistics structures and features and strategies. Teachers were instructed to use the criteria sheet and to practise before the actual assessment. They were asked to highlight the performance descriptors that matched the performance they observed; then decide on students' performance levels in a scale from 1 to 4. Teachers were also advised that seeing student performance shift from various levels across all aspects of language was normal. More details of the assessment criteria can be obtained from Appendix D.

3.5.4. *Data collection procedures*

Data collection was conducted via a three-hour accredited professional development workshop delivered and trained by an assessment specialist. Teachers signed up for either a morning session or an afternoon session. There were 17 teachers signed up for both sessions; however, five teachers later withdrew.

The workshop was delivered once in the morning and once in the afternoon following the same procedure. First, teachers checked in by signing an attendance sheet and providing their email address. The email address would become the main communication channel if they agreed to be followed up. After registration was completed, the research project was specifically described to the teachers. Teachers indicated their formal agreement to take part in the research by signing the participant statement and consent form (see Appendix A). After this, the trainer began the training section. She explained in detail about the TEAL project and how all the research materials were adapted from it. This was followed by detailed explanations of how the tasks had been designed and assessment criteria developed. Information on how to use the criteria was strongly emphasised. To illustrate, one common sample of student performance and scored criteria was provided. If teachers had any questions or comments, they were encouraged to ask or share these with the trainer. All teachers then watched and scored one student performance sample together. This gave the teachers a sense of what they would need to do in the actual assessment activity. No discussions were allowed during each assessment. After this, teachers were given a set of three different assessment sheets that were then used to score three different tasks. Before scoring each student performance sample, teachers were advised to read the criteria sheet carefully and clarify with the trainer anything they did not understand. They were then shown the video of each student sample twice. During the first time watching the first student sample, teachers were encouraged not to refer to the criteria; however, they could use the criteria sheet the second time.

When scoring they were asked to highlight performance indicators in each assessment strains in the criteria that they thought matched with the student's ability in a

4-point scale. In addition, they could add any comments they thought would justify and support their final decisions they made against the student.

Right after finishing scoring the first student performance sample, teachers were assigned to work in groups of three, with four groups in total. Discussion focused on the two guiding questions: ‘Compare your responses. What was similar and what was different? Why did you have differences?’ Teachers began the discussion by introducing themselves so that the researcher could recognise their voices when transcribing. With the teachers’ permission, all discussions were audiotaped. The recordings were then transcribed and analysed to design the interview question for the follow-up retrospective think-aloud activity and to triangulate with the information collected from the think-aloud activity.

3.5.5. Retrospective think-aloud activity

Think-aloud methods have been widely employed in previous studies in language assessment (Barkaoui, 2007; Cumming, 1990, 2002; Lumley, 2002b; Weigle, 1999). Although these studies were conducted in the context of writing assessments, they aimed to obtain insights into assessors’ socio-cognitive processes. Broadly, there are two types of think-aloud procedures: concurrent think-aloud (in which participants verbalise their thinking while they perform a task) and retrospective think-aloud (in which participants first perform the task in silence and then recall and verbalise their decisions). Retrospective think-aloud has been reported to help teachers give more verbalisation (Bowers & Snyder, 1990; Van Den Haak, De Jong, & Schellens, 2003). This study aimed to examine how teachers justified the assessment decisions they made, therefore, retrospective think-aloud protocols were more suitable as teachers

would then focus solely on the scoring task. Participants were asked to rate the video performances and verbalise their justification of their assessment decisions afterwards.

It is argued that think-aloud is a complex process and; therefore, not all people can do it effectively. However, research suggests there are more problems with concurrent than retrospective think-aloud (Van Den Haak, De Jong, & Schellens, 2003). As the teachers in this study first completed the assessment process and then justified judgement decisions that they had made, they were not asked to assess and justify at the same time. As this is a less complicated process compared with concurrent think-aloud protocol, training for retrospective think-aloud was really not a concern.

Apart from teachers' oral justification, interviewing was also employed. In this activity, interviews were conducted in the form of a follow-up, immediately after the teachers' justification of their assessment decisions. The follow-up interviews were semi-structured interviews that included a set of primary questions. The interview questions were divided into three major categories to discover information about the teachers' confidence in assessment, their process of assessment and their assessment biases (see Appendix E). Semi-structured interviews were conducted with individual teachers after they completed their think-aloud. Semi-structured interviews were chosen as a follow-up to the teachers' justifications, as these offered the opportunity to obtain more comprehensive data. In addition, interviewing was conducted systematically. The researcher as interviewer introduced himself to the interviewed teachers as a fellow teacher, but an outsider, from a foreign English assessment context. Research shows that outsiders are perceived by interviewees as neutral and therefore are given more information as compared to insiders (Fonow and Cook, 1991) . Therefore, the positionality and power of the researcher in this study was not believed to have any effect on teachers' responses. Additional questions enable participants to elaborate their

responses to the predetermined questions and, along with the flexibility of the interviewer, enable conversations to flow naturally. Furthermore, such interviews tended to become situational and conversational, allowing the researcher to gain a richer and deeper understanding of the characteristics of assessment practice behaviour of those teachers who gave inconsistent assessments. An interview guide was developed consisting of predetermined open-ended questions aimed at asking the participants to justify and explain their decisions (see Appendix E). All individual interviews were audiotaped with the consent of the participants.

It is not surprising that the interview is the most widely used method in qualitative research studies. The interviewer who is the researcher or someone who is working for the researcher collects data by asking the interviewees (the research participants) several predetermined questions in a highly structured style, follow-up questions and an opened-ended style (Johnson & Christensen, 2000). Qualitative interviews, as suggested by Johnson and Christensen, enable the researcher to 'enter into the inner world of another person and to gain an understanding of that person's perspective' (p. 144). The approach to qualitative interviewing used in this study is the interview guide approach. One of the characteristics of the interview guide approach asserted by Johnson and Christensen is that the interviewer predetermines and outlines the topics and questions to be discussed. The questions may be conducted in any specific order and are reformulated and reworded as the interview proceeds. This approach is administered in a relatively structured style—the interviewer must keep the interview on track by asking all interviewees the same set of broad questions, but not necessarily in the same order.

Interviews are an excellent way to gather detailed information. Whatever topic is of interest to the researcher employing this method can be explored in much more depth

than with almost any other method. Qualitative interviews are to elicit detailed and in-depth information, they are especially useful when a researcher's aim is to study social processes, or the "how" of various phenomena. In addition, another benefit of interview guided approach is that researchers can make observations beyond those that a respondent is orally reporting. A respondent's body language, and even her or his choice of time and location for the interview, might provide a researcher with useful data.

However, qualitative interviews rely on respondents' ability to accurately and honestly recall whatever details about their lives, circumstances, thoughts, opinions, or behaviours that are being asked about. According to Esterberg (2002), observation should be more effective than interview for those who want to know about what people actually do, rather than what they say they do. In addition, success of research interview may also a lot depend on the interviewer's ability, especially in semi-structured interviews because face-to-face interviews are characterised by synchronous communication in time and place (Opdenakker, 2006).

Johnson and Christensen (2000) note that conducting a qualitative interview successfully will establish rapport with the participants. However, impartiality or neutrality is needed in the interviewer responses, as failure to do so may result in bias. A positive or negative reaction to participants' responses might affect further information they are about to provide. Probes—prompts for clarifying responses and adding information—should be used to increase effectiveness.

3.6. Stage 2

3.6.1. Participants

All participants in the first stage were followed up in the second stage, as all teachers taking part in the first stage of this study agreed to be contacted for further participation. The second stage consisted of a retrospective think-aloud protocol and interviews.

3.6.2. Retrospective think-aloud

As stated earlier, the purpose of this activity was to collect teacher justifications of their assessment decisions and to gain more insights on the process by which their decisions were made. Teachers were contacted by email to arrange a time and venue to for the think-aloud.

Prior to the meeting with the researcher, the participant was advised not to prepare anything because they would have a chance to view their work again. First, the researcher explained in detail the purpose of the think-aloud, what the teacher was expected to do and the risks they might face when taking part in the activity. By signing the consent form, the participant agreed to participate in the second stage of the research and agreed to be audiotaped. Following this, the participant was provided with their scored criteria sheet that they had generated in the first stage. At the same time, they had to watch the same videos they viewed earlier. They were asked to review the video of each task and again observe the scored criteria sheet of that task. They were advised not to make any amendments or changes to the original score sheet, but they could take notes on a separate piece of paper. The participant then did the same thing when reviewing videos of the other two tasks. After the participant had finished reviewing the

videos and the scored criteria sheets again, they were asked to justify the judgement decisions they had made. Specifically, they explained to the researcher why they had assigned each student at certain levels of performance. There was no time limit for this session.

Once the participant had finished verbalising, the researcher asked them a few questions regarding their assessment process. The follow-up interview questions were focused on three dimensions of the assessment process. The first dimension was to examine the level of confidence in teacher assessment. In particular, the participant was asked to indicate how confident they were when they assessed the students and after that specifically explain what accounted for their confidence. The second dimension focused on the process of assessment. The teachers were asked whether they referred to the assessment criteria sheet when they first watched the video. This question was to identify the initial elements of student oral language development they looked for to make their assessment decisions and whether examining the criteria changed their first thoughts regarding student level. In other words, it was important to understand if there were any differences in their assessment decision-making when they used the criteria and when they did not use the criteria. It was also important to see critical the teachers' first impressions of student performance were in their decision-making. To gain this kind of data, the participant was asked about their first impression of the strengths of each student. Then, they had to indicate whether their first impression had any effect on their final assessment decision and justify their answer. Finally, in the process-focused dimension, the participant was asked whether, during the scoring process, they looked for plus points in student performance and gave them a higher score or looked for minus points to give them a lower score. In the final dimension, attention was drawn to identifying factors that influenced teachers' assessment decisions, including student

language background, gender, teacher qualification, related teaching experience and task genres. For this study, with the participants' agreement, all talks and interviews were audiotaped.

3.7. Pilot Study

Prior to its actual administration, the assessment activity was piloted by five teachers who were teaching and working with EAL/D students at the time the pilot was conducted. However, the pilot was conducted online while the actual assessment process was conducted on a face-to-face basis. This was because the recruitment site changed from Victoria to NSW. The only differences between the online administration and face-to-face administration were the training package and versions of assessment materials. In the pilot, due to the distance issue the training section was provided to the participants in electronic form including an assessment guide and sample scored criteria sheets. However, in the actual assessment activity training was conducted by a language assessment expert. The face-to-face training was more effective because the participants could ask questions and clarify their understanding with the trainer to ensure they would complete the assessment activity as it was designed. Another difference was that teachers worked on the soft copy of the criteria, on which they could highlight the descriptors they chose. However, in the hard copy version teachers were asked to underline or highlight the performance descriptors that applied to student performance level. Overall, these differences did not appear to have any influence on how teachers scored.

3.8. Data Analysis

3.8.1. Analysis of assessment data

Since the purpose of analysing the teacher scores was to obtain information on teacher variability and consistency, calculations were conducted to find out the mean scores. As explained earlier, 12 teachers agreed to take part in the research study. Each teacher marked three student outputs using the criteria including seven assessment categories. Individual marks are taken as separate subsamples for data analysis. In other words, the individual judgment of teachers in each category was considered as a distinct variable; therefore, each teacher assigned 21 scores, making up for 252 observations (See example in Table 4.3). These can be traced back to the number of teachers. This number of observations was large enough for the purpose of analysis. However, given this was still a fairly simple data set, all data collected from the assessment activity were manually calculated. For the purpose of calculation, data were first modified prior to primary analyses being conducted.

3.8.1.1. Modification for variability analysis

All the teachers were asked to watch the student performance, highlight all the performance descriptors in the criteria that they thought would match with the student language capability and decide at which overall level they would place that student. The overall levels were then treated as scores they assigned to each student. Only one score, a final score, was produced for each student. In addition, participants were reminded that their final decision had to be based on their highlighted performance descriptors.

However, as commonly seen in the literature, it was necessary to obtain information on the macro-levels of each overall score to examine the teachers' tendency

in terms of stringency and consistency when assessing various aspects of student performance. As previously described, there were seven categories in the criteria consisting of communication, cultural conventions, text structures, grammatical features, vocabulary, phonology and strategies. Although the teachers were not asked to give sub-scores for individual aspects of the criteria, some of them did. Interestingly, most of the teachers who did not assign sub-scores did so indirectly. Their highlighted performance descriptors had already told the story. Their sub-scores were generated based on where the highlighted descriptors were.

There were still some cases in which, for example, in one assessment category of the criteria, performance indicators that were highlighted were scattered across all levels. In such cases, the sub-scores for those categories were decided by an analysis of the teachers' explanations during group discussions following completion of the assessment of each student sample. Some teachers whose sub-scores were not identified in their assessment sheets shared their decisions across the assessment categories with other teachers in their group. Thus, instead of 12 final scores from all the teachers each student was also assigned another 84 sub-scores across seven performance areas. That is one student received 7 sub-scores from each teacher, so 12 teachers gave each student a total of 84 sub-scores.

The first aim of this study was to determine how consistent teachers' assessments were. However, an analysis of consistency in teachers' assessment was dependent on the variability levels of such assessment. Once data on variability were determined, further analysis was conducted to identify trends in relation to consistency.

To facilitate variability calculation, it was necessary to provide a better and clearer understanding of variability in the different dimensions of the data set.

Therefore, prior to conducting any calculations, explanations of variability were defined

along three different dimensions (e.g., variability for individual students, variability across students and variability for individual assessment categories). Variability for individual students can be understood as a score (an observed score) assigned by a teacher to a student that is greater or less than the mean score assigned for that student. If the difference between the observed score is greater than the mean score, it means that teacher was more lenient than the other participating teachers, but it does not mean that she was more lenient compared with what the student was supposed to receive. The reverse means her assessment for that student was stringent compared with those of the other teachers. Similarly, variability across students is the case in which an observed score by a teacher is greater or less than the mean score across three students, one from each task. In other words, variability in this dimension means the difference between the mean score and the observed score. Finally, the explanation for variability for individual assessment categories is that an observed score by a teacher is greater or less than the mean score for a certain category across the students.

To determine whether an individual teacher's assessments were lenient or stringent, single calculations had to be conducted on scorings assigned for individual student performances. For instance, there were three student performances and 12 teachers. To examine whether a teacher was more lenient or more stringent compared with other teachers, it was problematic to include scorings for all performances in one calculation because the mean score and deviation were different for each performance. Therefore, calculations were first conducted on each performance. There were two ways of calculating teacher variability and these are presented below. However, only one calculation was used and the description of the other one is aimed to support and justify the chosen approach.

The first approach is presented through the formula as follows:

$$\frac{\sum(x-\bar{x})}{n}$$

For each performance, one score from each teacher was used at a time. To illustrate, a calculation of variability on communication of Student 1 (S1) was demonstrated in Table 3.2. As can be seen in Table 3.2, the 12 teachers in the first column gave 12 different sub-scores in the second column to S1 on communication. The scores were then added up and divided by 12 to acquire the mean (2.83) that was then subtracted from individual scores. Whether a teacher was lenient or not was decided based on how big the difference between their score and the mean score. The greater their score was, compared to the mean, the more lenient (or less stringent) their assessment on this performance area was and vice versa. Ideally no difference was expected in the assessment. Table 3.2 shows that teachers T01 and T12 produced the most lenient assessments on communication for S1 at 1.17, whereas the least lenient or most stringent assessments were assigned by teachers T03, T05, T06 and T11.

Table 3.2

Non-Scaled Deviations on Communication for S1

Teacher	Score	Deviation
T01	4	1.17
T02	3	0.17
T03	2	-0.83
T04	3	0.17
T05	2	-0.83
T06	2	-0.83
T07	3	0.17
T08	3	0.17
T09	3	0.17
T10	3	0.17

T11	2	-0.83
T12	4	1.17
Mean	2.83	

Note. T: teacher, S: student

Absolute deviations are the basic measurement units that were used for all variability and consistency throughout this study. However, in calculating variability actual absolute deviations were not used for two reasons. The first reason is that it would be invalid if the actual absolute deviations were added to make the mean absolute deviations, due to the differences in mean scores across students and categories resulting in different values in absolute deviations. Let us consider variability for communication across students as an example. As can be seen in Table 3.3, the mean scores on communication across all three students are different and differences are also found in absolute deviations. Therefore, the mean absolute deviations cannot be used to decide if a teacher assessment is lenient or stringent. Any conclusions drawn on these are invalid and often questioned.

Table 3.3

Actual Absolute Deviations Across Students

Teacher	S1	S2	S3
T01	1.17	1	-0.25
T02	0.17	0	-0.25
T03	-0.83	1	0.75
T04	0.17	0	0.75
T05	-0.83	1	0.75
T06	-0.83	0	-0.25
T07	0.17	0	-0.25
T08	0.17	-1	-0.25
T09	0.17	-1	0.75

T10	0.17	0	-0.25
T11	-0.83	-1	-1.25
T12	1.17	0	-0.25
Total	2.83	3.00	3.25

Note. T: teacher, S: student

Due to the invalidity and questionability of the first calculating method, another way of calculation was adopted. The only difference between the two methods is what is called scaling. To cope with the issues of invalidity and questionability, scaling was applied on the absolute deviations. In

Table 3.4, there were three values in 12 absolute deviations for each student (e.g., -0.83, 0.17 and 1.17 for S1, -1, 0 and 1 for S2 and -1.25, -0.25 and 0.75 for S3). These values were then levelled up. For instance, for S1 the smallest parameter was numbered 1, the second smallest was numbered 2 and the biggest was numbered 3. There were two cases in which there were more than three parameters for each student. In these instances, the same procedures were repeated until the biggest was reached. Details and instances of scaling are exemplified in Table 3.4.

In this table, the first column contains identity codes for teacher from Teacher 1 (T01) to Teacher 12 (T12). Information for three students S1, S2, and S3 is presented in the last three columns. Each is divided into two sub-columns, one for the actual deviations and the other for scaled deviations. Through scaling, the difference among the mean scores was no longer a concern threatening the validity of calculation. All calculations were conducted using scaled deviations.

Table 3.4

Scaled Absolute Deviations

Teacher	S1		S2		S3	
	Deviation	Scaling	Deviation	Scaling	Deviation	Scaling
T01	1.17	3	1	3	-0.25	2
T02	0.17	2	0	2	-0.25	2
T03	-0.83	1	1	3	0.75	3
T04	0.17	2	0	2	0.75	3
T05	-0.83	1	1	3	0.75	3
T06	-0.83	1	0	2	-0.25	2
T07	0.17	2	0	2	-0.25	2
T08	0.17	2	-1	1	-0.25	2
T09	0.17	2	-1	1	0.75	3
T10	0.17	2	0	2	-0.25	2
T11	-0.83	1	-1	1	-1.25	1
T12	1.17	3	0	2	-0.25	2

Note. T: teacher, S: student

3.8.1.2. Modification for consistency analysis

One of the primary aims of this study was to examine how consistent teacher assessments were when they scored student oral language performances. Findings on teacher consistency shed the light on the process of teacher assessment decision-making about students' oral language development. In this section, explanations and justifications of analyses conducted to examine consistency of teacher assessments are first presented. Analyses for consistency were conducted along three different dimensions (e.g., consistency for individual students, consistency across students and consistency across categories). Consistency for individual students is the extent to which an observed score given by a teacher to a student is close to the mean score.

Table 3.4

Scaled Absolute Deviations

Teacher	S1		S2		S3	
	Deviation	Scaling	Deviation	Scaling	Deviation	Scaling
T01	1.17	3	1	3	-0.25	2
T02	0.17	2	0	2	-0.25	2
T03	-0.83	1	1	3	0.75	3
T04	0.17	2	0	2	0.75	3
T05	-0.83	1	1	3	0.75	3
T06	-0.83	1	0	2	-0.25	2
T07	0.17	2	0	2	-0.25	2
T08	0.17	2	-1	1	-0.25	2
T09	0.17	2	-1	1	0.75	3
T10	0.17	2	0	2	-0.25	2
T11	-0.83	1	-1	1	-1.25	1
T12	1.17	3	0	2	-0.25	2

Note. T: teacher, S: student

3.8.1.2. Modification for consistency analysis

One of the primary aims of this study was to examine how consistent teacher assessments were when they scored student oral language performances. Findings on teacher consistency shed the light on the process of teacher assessment decision-making about students' oral language development. In this section, explanations and justifications of analyses conducted to examine consistency of teacher assessments are first presented. Analyses for consistency were conducted along three different dimensions (e.g., consistency for individual students, consistency across students and consistency across categories). Consistency for individual students is the extent to which an observed score given by a teacher to a student is close to the mean score.

Consistency across students is the extent to which an observed score by a teacher is close to the mean score consistently across students. Consistency across categories is the extent to which observed scores by a teacher across students are close to the mean score consistently across categories. While the actual absolute deviations were used to analyse teacher variability, scaled absolute deviations were used instead to avoid invalidity and unreliability of the findings. Similarly, again the actual absolute deviations were not utilised in analyses of consistency.

The primary purpose of examining consistency is to examine the distance between the observed score and the mean score—how close or far the distance is. It is not necessarily a concern if the observed score is greater or less than the mean score. Therefore, all negative absolute deviations (i.e., less than the mean) were modified and treated as positive values when analysing consistency. Details of modification of negative absolute deviations are exemplified in Table 3.5. This is an example of teacher scorings for S1 on cultural conventions. This table illustrates that six out of 12 teachers have smaller scores than the mean score, namely -0.67 . Therefore, these teachers assigned stringent scores to this student on this language area. However, the main aim is to explore consistency, not variability. How far the deviations are from the mean is really the concern, not how big the difference is in terms of value.

Table 3.5

Sampled Modification of Absolute Deviations

Teacher	Observed Score	Actual Deviation	Modified Deviation
T01	4.0	1.33	1.33
T02	4.0	1.33	1.33
T03	2.0	-0.67	0.67
T04	2.0	-0.67	0.67
T05	2.0	-0.67	0.67

T06	2.0	-0.67	0.67
T07	3.0	0.33	0.33
T08	3.0	0.33	0.33
T09	3.0	0.33	0.33
T10	2.0	-0.67	0.67
T11	2.0	-0.67	0.67
T12	3.0	0.33	0.33
Mean	2.67		

Note. T: teacher, S: student

3.8.2. Analysis of questionnaire data

Demographic information collected from 12 participants through the questionnaire was first cleaned to ensure its quality (Dörnyei, 2007). Responses from close-ended questions were turned into numerical data and analysed using descriptive statistics methods through the statistical computer software SPSS. The questionnaire data were then analysed in conjunction with the assessment data. Findings from these analyses were triangulated with the information obtained from the think-aloud protocols to answer the second research question.

3.8.3. Analysis of document data

To support analyses exploring the effects of student-related variables and task-related variables, the task annotations and videos from the TEAL website were consulted as part of the data analysis. Two pieces of information were included in teachers' annotations of tasks. The first piece included the information on task types and task genres, while the second piece contained the data on students' background information such as gender and backgrounds. Videos were also consulted to provide the data on students' accent.

3.8.4. Analysis of interviews and group discussion data

The primary purpose of the think-aloud protocol, interviews and discussions was two-fold. First, data from these sources was used to identify whether teacher assessment behaviours were driven by any internal and external factors during the assessment process. Second, this data was also used to identify characteristics of extreme assessment behaviours, that is, the most and least lenient and the most and least consistent. As the first step of this analysis, the think-aloud and interviews as well as the group discussions were transcribed. The coding scheme suggested by Cumming, Kantor, and Powers (2002) was adopted to identify influential factors, with data coded by both relying on predetermined themes identified through the literature and by using grounded theory (Glaser & Strauss, 1967). The grounded theory coding approach was mainly used to untangle concerns about the characteristics of extreme assessment behaviours. In addition, to facilitate the coding process, a computer program NVivo version 10 was used to help the researcher manage the work of coding.

3.9. Ethical Considerations and Research Validity

This study involves humans as participants; therefore, approval for research ethics was needed (Punch, 2009).. This study considered all potential ethical issues that may have arisen not only before, but also during and after conducting the research (Miles & Huberman, 1994). Ethical issues included deception, obtaining participant consent, freedom to withdraw, confidentiality and anonymity of participants (Johnson & Christensen, 2000) and the use of data as the researcher intended. Ethical approval was sought from the University of New South Wales (UNSW) only. Ethical approval from the NSW State Education Research Approvals Process (SERAP) was not required as the teacher participants took part in this research on Saturdays at UNSW as part of their

own professional development, and only assessment material available publicly on a website designed for teacher assessment training were used.

In terms of research validity, as Onwuegbuzie and Johnson (2006) note, mixed methods research integrates strengths and weaknesses of quantitative and qualitative research; therefore, assessing its validity involves complexity. At some stage of mixed methods research, a researcher includes themselves as a human research instrument that collects and interprets data. This may create one type of ‘threat to validity’, called researcher bias (Johnson & Onwuegbuzie, 2004; Onwuegbuzie & Johnson, 2006). That is, the researcher may report the results based on what they want to discover (Johnson & Christensen, 2010). Thus, data collection and interpretation may be skewed by the subjectivity of the researcher. To avoid this bias, I adopted a stance of ‘reflexibility’ or self-reflection, rigorously monitoring my own judgement so that less bias is introduced in the data interpretation process, and continually checking my interpretations with my supervisors and peers.

Chapter 4. Variation in Teacher Scorings

4.1. Introduction

In the previous chapter on methodology, a mixed methods research approach with more emphasis on the qualitative component and the design of this study was described in detail, followed by a full description of how the study tools were used and data were collected. This chapter first reports the findings collected from the questionnaire and the teacher assessment activity. As reviewed in the literature presented in Chapter 2, variability exists among teachers as raters in large-scale testing contexts and as assessors in their classrooms. The teacher assessment activity described in Section 3.5.3. examined how consistent teachers were when they assessed students' oral language development. The activity was conducted to answer the first research question of the study: 'To what extent does the teacher-based assessment of students' oral English communication skills produce consistent results (i.e., across different teachers)?'. The literature review suggested that assessments made by teachers are influenced by several factors related to teachers themselves (e.g., their age, teaching position, experience and qualifications). In this study, such information as well as their background information were collected through the questionnaire described in Section 3.5.2. and other factors related to students and tasks also compiled. Findings on these factors are presented and discussed later in Chapter 5.

In terms of terminology, the terms *harshness* or *severity* are too psychometric in nature, and thus, do not suit classroom-based assessment contexts. It is basically unfair to conclude whether a teacher is harsh or lenient without taking assessment context into consideration. These terms create unpleasant feelings among teachers who are being

judged. For this reason, the term *stringency* is used in my thesis to indicate the extent to which teachers give low scores to student performances. It is understood that a teacher with stringent assessment has strict requirements about student performance and this teacher tends to take into account contextual factors when making assessment decisions. In this study, the term *stringency* is sometimes used interchangeably with *variability*.

4.2. Variations for Individual Students

As explained earlier, variability is observed when a score given by a teacher to a student is greater or smaller than the mean score. The calculation of this dimension of variability is performed by

$$\frac{\sum(scale(\chi - \bar{\chi}))}{n}$$

where \sum represents summation, χ represents the observed score, $\bar{\chi}$ represents the mean score and n represents the number of observed scores across categories. The higher the results are, the bigger the gap between the observed score and the mean score. This means more variation is observed. As shown in Table 4.1, there are differences in the degree of variation among teachers in scoring S1's performance. Notably, the highest value at 2.86 is found for teacher T02 followed by teachers T01 and T09 at 2.57 and 2.43, respectively, meaning that these teachers gave the most lenient assessments to this student. Whereas, teachers T04, T10 and T11 are found at the other end of variability continuum at 1.14, indicating that they assigned the least lenient assessment to this student. It can also be seen from the most and the least lenient assessments that the range of variability degree is quite large at 1.72, indicating a considerable difference was observed among those classroom assessors.

Table 4.1

Variability for Individual Students

Teacher	S1	S2	S3	Overall
T01	2.57	1.71	1.86	2.05
T02	2.86	2.29	1.43	2.19
T03	1.86	2.71	2.29	2.29
T04	1.14	2.43	1.71	1.76
T05	1.57	2.00	1.71	1.76
T06	1.71	1.86	1.57	1.71
T07	1.57	2.29	1.71	1.86
T08	1.57	1.29	1.86	1.57
T09	2.43	1.71	1.86	2.00
T10	1.14	2.29	1.57	1.67
T11	1.14	1.57	1.43	1.38
T12	1.71	1.29	1.00	1.33
Mean	1.77	1.95	1.67	
Range	1.72	1.42	1.29	

Note. T: teacher, S: student

Further, Table 4.1 also shows that the range of 1.42 between teachers with the least lenient and the most lenient assessments suggests noticeable differences among teachers when they assessed performances made by S2. Specifically, the highest degree of variability belongs to T03 at 2.71 followed by T04, T07 and T10 at 2.43 and 2.29, respectively, meaning that this S2 received the highest score from these teachers. Conversely, teachers T08 and T12 are identified as giving the lowest score to this student when the variability degree is found at 1.29, meaning they stringently scored the performance by S2.

Finally, it can also be observed from Table 4.1 that teacher T03 again has the highest degree of variability at 2.29 when they assessed the performance of S3. The

second highest position at 1.86 is found for teachers T01, T08 and T09 followed by teachers T04, T05 and T07 at 1.71. Like the case of T03, teacher T12 is found a second time at the bottom of the variability continuum at 1.00. The distance between the highest and lowest degree of variability is 1.29, also suggesting that the discrepancies among the classroom assessors should not be ignored to any extent.

Consistency for individual students is the extent to which an observed score by a teacher for a student is close to the mean score across categories. Variables that were used to calculate this dimension of consistency are χ which is the observed score, $\bar{\chi}$ the mean score and n the number of observed scores (categories). The formula for this is

$$\frac{\sum(\chi - \bar{\chi})}{n}$$

As a reminder, the negative absolute deviations were changed to positive. The results presented in the last row of

Table 4.2 show that, in general, given that the smaller the distance is the more consistent the assessments are, teacher assessments for S3 are the most consistent compared with their assessment for the other two students. Specifically, attention is drawn to the mean consistency across teachers for S3 of 0.45, meaning that not many variations were found among teacher scorings for this student. Data in the fourth column show that the gap between the least consistent assessment and the most consistent assessment is 0.26.

Meanwhile, being the least consistent with the mean consistency at 0.69, teacher assessments for S1 exhibit a wider gap in their assessments ranging from 0.43 to 1.12. This finding suggests that teachers may have differed from each other in understanding the performance of this student, interpreting the assessment criteria for this task and probably weighing different language areas. Finally, assessments for S2, in terms of consistency, are found between S3 and S1, but a little closer to the S1 with the

variations ranging within 0.60, meaning that teacher differences in scoring this student should be taken into consideration to some extent.

Table 4.2

Consistency for Students

Teacher	S1	S2	S3	Mean
T01	1.10	0.63	0.32	0.68
T02	1.12	0.70	0.46	0.76
T03	0.55	0.68	0.58	0.60
T04	0.64	0.42	0.54	0.53
T05	0.76	0.87	0.39	0.67
T06	0.43	0.27	0.42	0.37
T07	0.57	0.49	0.51	0.52
T08	0.43	0.82	0.42	0.56
T09	0.69	0.51	0.39	0.53
T10	0.64	0.47	0.32	0.48
T11	0.67	0.70	0.51	0.63
T12	0.74	0.75	0.56	0.68
Mean	0.69	0.61	0.45	
Range	0.69	0.60	0.26	

Note. T: teacher, S: student

4.3. Variations of Individual Teachers Across Students

An explanation to clarify understanding is necessarily an important reminder for the researcher prior to conducting any calculations relating to this dimension of variability. Variability across students is when an observed score given by a teacher is greater or less than the mean score across all students. To support understanding on how the calculation was conducted, a formula is provided as follows:

$$\frac{\sum(scale(x - \bar{x}))}{n}$$

$$N$$

In this formula, N represents the number of cases. Other variables that were used to calculate this dimension of consistency are χ which is the observed score, $\bar{\chi}$ the mean score and n the number of observed scores (categories). The overall results are detailed in Table 4.1. Statistics in this table shows that 11 out of 12 teachers differ from each other in the degree of variability, ranging between 2.29 and 1.33. Notably, teacher T03 is at the top end of the variability continuum, closely followed by teacher T02 at 2.19 and then by teachers T09 at 2.05 and 2.00, respectively, meaning their assessments for all the students were the most lenient. At the other end of the continuum is the presence of teachers T12 and T11 whose degree of variability was determined to be 1.33 and 1.38, respectively, indicating that they assigned the most stringent scores to this student. With the degree of variability at 1.86, 1.71, 1.67 and 1.57. respectively. teachers T07, T06, T10 and T08 are found in the middle range of the continuum. Positioned at 1.76, teachers T04 and T05 contribute to a relatively even distribution on the continuum of variability.

An exploration as to whether teachers' variability remained stable or varied when they assessed individual student performances shed light on explaining why the teachers exhibited certain assessment behaviours. This exploration did not aim to examine which teacher has the highest degree of variability in their assessment or who made the least lenient assessment. It was only to see if a certain teacher's variability level was stable when they assessed three different students. Such observations were made by examining the difference between the teacher's highest and lowest degrees of variability across three cases. As presented in Figure 4.1, there are two opposite but noticeable tendencies in terms of variability. In the one tendency, teacher T06 has the least variations in her degree of variability within a range of 0.29 scoring performances by the students. This suggests, regardless of whether they are lenient or stringent, that

this teacher may assign stable scores when they are asked to assess students' oral language performances. Very few variations are also found in the cases of T11 and T05 whose variability was identified to be less than 0.5 degrees. In the other tendency, it is noticeable that teachers T02 and T04 have the most variations in their variability levels, even though they are not necessarily the extreme assessors in terms of variability. The gap between their highest and lowest degrees of variability is slightly less than 1.5, indicating that whether their assessment of student oral performances is lenient or stringent is unpredictable. As also seen in Figure 4.1, considerable variations are counted for teachers T01, T03 and T09 with the range of slightly over 1.00, meaning that their variability degree is not easy to predict to some extent.

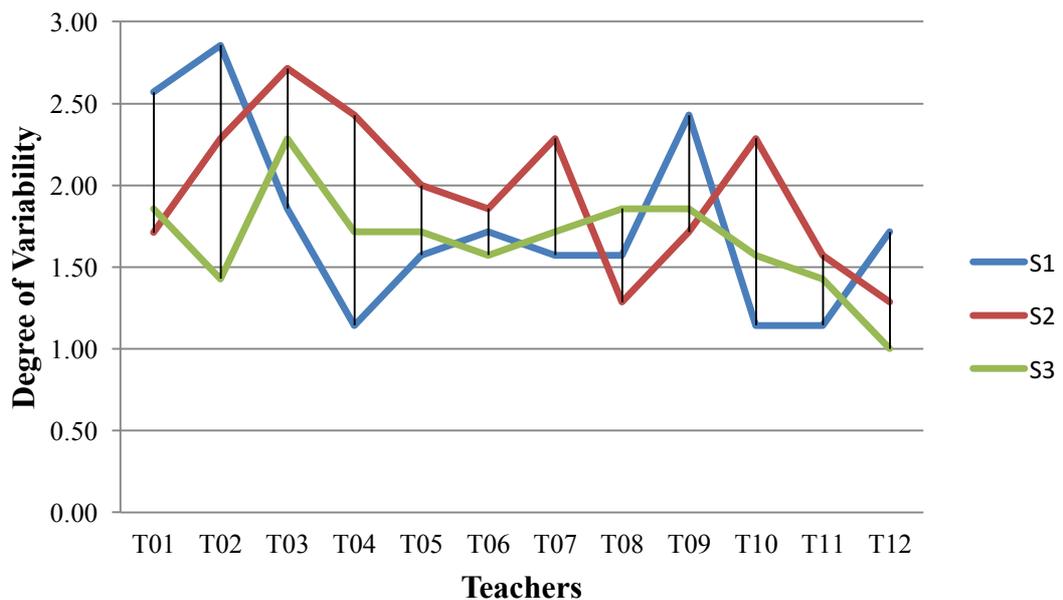


Figure 4.1. Variability across Students.

It is also shown in Figure 4.1 that differences in variability among teachers are found in the assessments they made of the first two students S1 and S2. In particular, the range of 1.72 teacher variability for S1 fluctuates from 2.86 to 1.14, implying that the result for this student may be questionable and, thus, may not be solely used to decide

this student's language ability. In the same way, it is also shown in Figure 4.1 that teachers demonstrated a great deal of variation in their assessments regarding how lenient they were in assessing S2's performance. With the degrees of variability ranging between 2.71 and 1.29 variability discrepancies for S2, together with the ones for S1, may be accounted for most of the overall variations.

Consistency across students is the extent to which an observed score by a teacher is close to the mean score consistently across students. The formula for calculating this dimension of consistency is as follows:

$$\frac{\frac{\sum(\chi - \bar{\chi})}{n}}{N}$$

In this formula, χ is the observed score, $\bar{\chi}$ the mean score and n the number of observed scores (teachers) and N is the number of cases. The results can also be seen in the last column of Table 4.1. This demonstrates that, in terms of teacher overall consistency or consistency across students, assessments by T02 are found to be the least consistent at 0.76, indicating this teacher did the assessment task considerably differently from the others. This finding also suggests it may be difficult and untrustworthy to make any conclusions or final decisions or even some generalisations about students' ability to use the language if solely relying on this teacher's assessments. Other teachers, T01, T12 and T05, may share T02's assessment characteristics based on their relatively low levels of consistency. Notably, 0.68 is reported for T01 and T12, closely followed by T05 at 0.67. Conversely, assessments made by teacher T06 are recorded as the most consistent and; therefore, may be counted on when making decisions on performances. Assessments by the other teachers are also described in this table as scattering over the continuum. Assessments with the highest

degrees of variability and most variations such as T02 and T06 will be further introduced and discussed in the following chapters.

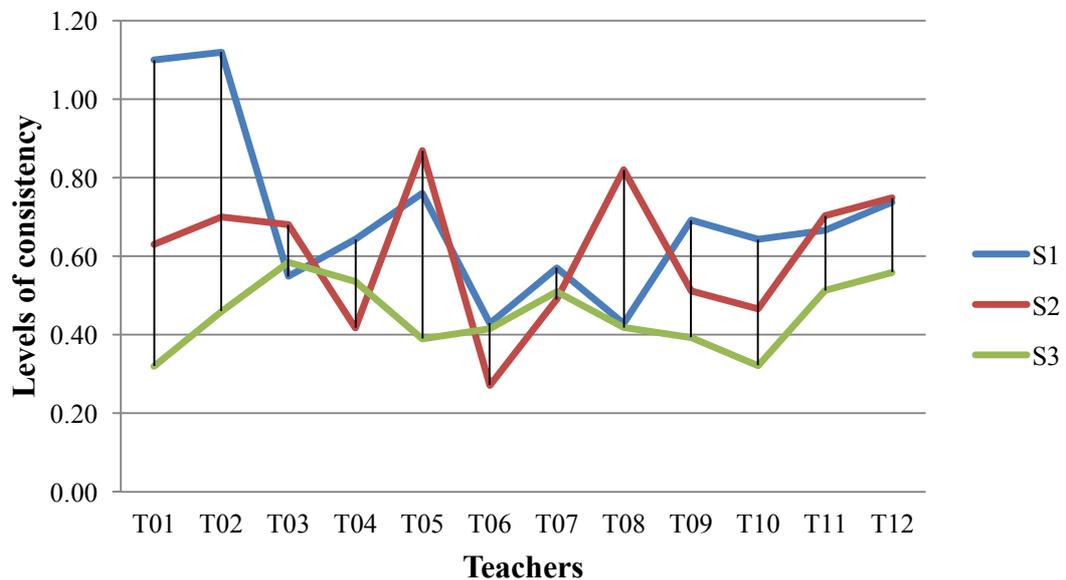


Figure 4.2. Consistency Stability across Students.

Apart from the information on the overall consistency of teacher assessments, it is also necessary to obtain more detail about how stable teacher judgements were across all the students. Information on consistency stability of teacher to some extent assists understanding and explaining overall consistency in their assessment practices. As presented in Figure 4.2, T07's assessments are not the most consistent, but are the most stable with a variation through three scorings of 0.06. This means the consistency level of the assessments by this teacher remained almost unchanged when they were scoring those students regardless of the students' individual differences. Similarly, assessments made by teachers T03, T06, T11 and T12 are also shown to exhibit modest amounts of variation, meaning that no matter how high their consistency levels were compared with other teachers, their assessment practice did not vary much over three assessment rounds. On the other hand, it is noticeable from Figure 4.2 that assessments by teacher

T01 and T02 kept changing dramatically across students. Specifically, with the lowest observed level of consistency at 1.10 for S1 and the highest at 0.32 for S3, the overall consistency of assessment by T01 was unpredictable. Whether the teacher was consistent or not in scoring one student did not appear to determine their consistency in scoring another student. Similarly, but more interestingly, it has been reported above that assessments made by T02 were the least consistent and were found to vary considerably and to be unpredictable. In large-scale operational testing situations, raters such as this teacher are required to attend retraining and may be excluded from assessment processes if no improvement is made after retraining (McNamara, 2001). However, in assessment for learning it is important to obtain more insights from this teacher to understand the entire process of what they considered, how they understood student performances and assessment rubrics and other factors of their assessments. Such investigations are presented and discussed in Chapter 5.

4.4. Variations for Individual Categories

Again, repeating an explanation of this dimension of variability is important. Variability for individual categories is the extent to which an observed score is greater or less than the mean score for each category across students. The scores used in calculations for this dimension were modified scores. However, raw scores that each teacher have to each student in all seven categories can be seen in Appendix F. Similar to calculations for the other two dimensions, scaling was applied on the absolute deviations. The formula for calculation this dimension of variability is as follow:

$$\frac{\frac{\sum(\text{scale}(\chi - \bar{\chi}))}{n}}{N}$$

This formula is similar to the one used for calculating variability across students; however, there is a significant difference. That is, in this formula, n represents the number of observed scores and it is the number of teachers in this case, whereas in the former calculation the n represents the number of categories.

Table 4.3

Variability for Individual Categories

Teacher	Com	Cul	Text	Gra	Voc	Pho	Stra	Mean
T01	2.67	1.67	2.00	1.33	2.00	2.00	2.67	2.05
T02	2.00	2.00	2.00	2.00	2.00	2.67	2.67	2.19
T03	2.33	2.00	2.00	2.67	2.00	2.33	2.67	2.29
T04	2.33	1.67	1.67	1.67	1.33	1.33	2.33	1.76
T05	2.33	1.33	1.00	1.67	1.67	2.00	2.33	1.76
T06	2.00	1.33	1.00	1.67	1.67	2.00	2.33	1.71
T07	2.00	2.00	1.67	1.67	1.33	2.00	2.33	1.86
T08	1.67	1.67	1.00	1.67	1.33	1.67	2.00	1.57
T09	2.00	1.67	2.00	1.67	1.67	2.33	2.67	2.00
T10	2.00	1.67	1.67	1.00	1.33	2.00	2.00	1.67
T11	1.00	1.00	1.33	1.33	1.33	1.67	2.00	1.38
T12	2.00	1.33	1.00	1.00	1.00	1.00	2.00	1.33
Mean	2.03	1.61	1.53	1.61	1.56	1.92	2.33	
Range	1.67	1.00	1.00	1.67	1.00	1.67	0.67	

Note. T: teacher, Com: Communication, Cul: Cultural Conventions, Text: Text Structures, Gram: Grammatical Features, Voc: Vocabulary, Pho: Phonology, Stra: Strategies

Table 4.3 shows the overall results of variability for each individual category in the assessment criteria across teachers and for individual teachers across categories. In terms of variability for categories across teachers, it can be seen from Table 4.3 that the degree of variability among teachers is highest for strategies and communication with

average at 2.33 and 2.03, respectively. This is followed by phonology with 1.92, meaning that teachers were lenient in their assessment for the students on these three assessment categories. Conversely, teachers are found to give less lenient assessments to students on text structures and vocabulary with the reported average degree of variability at 1.53 and 1.56, respectively. Further, the other two categories of cultural conventions and grammatical features are located at the same level at 1.61 in the variability scale. In the area of Strategies, the degrees of variability of the teachers are relatively close to each other when the gap between the most lenient assessment and the most stringent assessment is only 0.67. Meanwhile, the gap observed on communication is comparatively big, at 1.00 degree higher than that of strategies. Regarding categories having the lowest degrees of variability, the average variability degrees are the same at 1.00 across cultural conventions, text structures and vocabulary.

The last column in Table 4.3 provides information about the average degree of variability of teachers across categories. This information can also be seen in the last column of Table 4.1, thus, it is not further reported, and more attention is drawn on reporting variations of individual teachers across categories instead. It can be observed that from one end of the continuum teacher T01 demonstrates the most variations in their variability level with the gap between the highest and lowest variability degree of 1.34. Overall, this teacher tended to give their most stringent assessments to student performances on grammar. However, they were quite tolerant with students when they assessed students' abilities to communicate and use conversational strategies. Further, the same degree of variability at 2.00 in their assessments on text structure, vocabulary and phonology suggests that their assessment stringency on these language areas can be predicted by looking at their stringency on any one of these. Located at the other end of the continuum, the degree of variability of teachers T01, T02, T08 and T10 is reported

to range within 0.67, meaning that their assessments in terms of variability were relatively stable and relatively predictable. With the degrees of variability ranging within 1.00, more than 50 per cent of the teachers are found scattering within the middle of the continuum. Regarding predictability of teacher variability across categories, it is shown that teacher T02 is the most predictable because they indicated the same degree of variability in five out of seven categories, followed by the cases of teachers T08 and T12 with four out of seven categories sharing the same degree of variability. This suggests that, as mentioned above, knowing how stringent they are in one category can predict their stringency in the others.

In terms of consistency, consistency for individual categories is the extent to which observed scores by a teacher are close to the mean score across categories. The formula for analysis of this dimension of consistency is as follows:

$$\frac{\frac{\sum(\chi - \bar{\chi})}{n}}{N}$$

In this formula, the χ is the observed score, $\bar{\chi}$ the mean score and the n is the number of observed scores (categories) and the N is the number of cases. Again, the modified absolute deviations were used for calculation.

As shown in Table 4.4, there is not much difference among categories in terms of the overall consistency levels, with the highest at 0.54 belonging to text structures, closely followed by vocabulary and phonology at 0.55. The lowest level of consistency is reported for cultural conventions, communication and grammatical features at 0.64, 0.63 and 0.62, respectively, leaving strategies at 0.58. However, when examining the distance between the most and the least consistent assessments in each individual assessment category, remarkable discrepancies are observed. Notably, the smallest gap of 0.23 found for text structures suggests that regardless of whether teacher assessments

are consistent or not these assessments were quite close to one another on the consistency scale. Conversely, strategies accommodate the most variations among teachers in terms of consistency at 0.94, preceding grammatical features, communication and phonology at 0.78 and 0.72, respectively. This finding provides that teachers were not consistent with one another when assessing students on these language areas. The overall consistency of teachers across categories shown in the last column provides that, as also seen earlier in Table 4.3, teacher assessments vary within the range of 0.39, from 0.37 for T06 to 0.76 for T02. It is also found that teacher consistency scatters evenly across the scale.

Examination of teacher consistency stability across the assessment categories is described in Table 4.4. It is interesting to note the most consistent assessment is not actually the most stable, or the least consistent assessment is not necessarily the most variable one. For example, the least consistent assessment by T02 is ranked at the third position in terms of consistency stability with the range of 0.61 and the most consistent T06 is located at the fifth position with the range of 0.55. Whereas, being in the middle of the consistency continuum, assessments by T03 and T11 are reported to accommodate the most unstable consistency patterns with the range of 0.81 and 0.72, respectively. This suggests that in assessments by those two teachers in terms of consistency across all assessment categories, their judgement tended to be highly unpredictable. Thus, it is difficult to predict how they assessed student performances in one language area by simply observing what they did in another. Conversely, assessments made by the other teachers possessed a modest amount of variations through all categories. Specifically, the gaps between the most consistent category and the least consistent category of most teachers are closely similar to one another, namely from 0.45 to 0.61. This finding demonstrates that no matter how consistent assessments

made by those teachers were, their scoring practice remained much more stable, compared to T03 and T11 in the other end of the continuum.

Table 4.4

Consistency for Individual Categories

Teacher	Com	Cul	Text	Gram	Voc	Pho	Stra	Mean
T01	0.86	0.92	0.64	0.78	0.59	0.50	0.47	0.68
T02	1.03	0.42	0.78	0.56	0.97	0.78	0.80	0.76
T03	0.86	0.70	0.45	1.06	0.47	0.45	0.25	0.60
T04	0.31	0.70	0.50	0.56	0.31	0.55	0.80	0.53
T05	0.86	0.70	0.55	0.44	0.97	0.67	0.47	0.67
T06	0.36	0.42	0.55	0.28	0.69	0.17	0.14	0.37
T07	0.47	0.42	0.45	0.78	0.69	0.61	0.25	0.52
T08	0.47	0.64	0.55	0.56	0.47	0.33	0.86	0.56
T09	0.64	0.64	0.45	0.78	0.25	0.50	0.47	0.53
T10	0.14	0.70	0.50	0.61	0.31	0.61	0.47	0.48
T11	1.03	0.75	0.50	0.44	0.31	0.50	0.86	0.63
T12	0.47	0.64	0.55	0.61	0.53	0.89	1.08	0.68
Mean	0.63	0.64	0.54	0.62	0.55	0.55	0.58	

Note. T: teacher, Com: Communication, Cul: Cultural Conventions, Text: Text Structures, Gram: Grammatical Features, Voc: Vocabulary, Pho: Phonology, Stra: Strategies

4.5. Discussion

Variability and consistency are inseparable components of variability and information on teacher variability can shed light on a better understanding of teacher behaviour in terms of consistency. For these reasons, any discussion on variability should be carried out in a way that findings about one aspect support understanding about the other. Results on variability and consistency in this study have previously been presented in the three dimensions; therefore, these will be discussed in the same

way with the two teacher assessment conceptions discussed one after another in each of the dimensions.

In the first dimension, assessments of individual students, teachers did not display the same level of variability when scoring student performances. Teachers are identified as behaving slightly differently from one another in assessing individual student performances. Teacher assessment practice for each student identified through the gap between the most lenient and the least lenient scores were remarkably differential. Information on consistency in this dimension also provides evidence for and fortifies the argument. This first instance of teacher differences means that, when observing student performances, teachers might hold different views of what a satisfactory performance looks like, even though they were asked to strictly follow the assessing criteria. Such a difference may be attributable to teacher differences in perceiving student language performances. For example, there is a tendency in which assessors discrepantly place particular emphases on different performance areas (Vaughan, 1991). To some teachers, a good or successful oral language performance by a student is judged based on how well that student can use grammar, the accurateness of their pronunciation or the richness of their vocabulary. Conversely, other teachers underemphasise these language-related performance components and instead give more focus on the purpose of the speaking task—what the student is told to do.

Moreover, technically, teachers are bio-sociologically unique—they are individually different in thinking and acting. Saying this does not necessarily mean that they are free to perform differently in their assessing tasks. They have been instructed to do the tasks and are given the same source of assessment materials. Therefore, their assessments are supposed to be close to each other. Differences in scoring may also suggest that the assessors may have taken a range of factors into consideration during

their decision-making process (Brown, Iwashita, & McNamara, 2005). Given that students as humans like the teachers themselves are unique, the teachers may consider characteristics related to students other than their ability in the performance task. These considerations may have resulted in the differences in variability and consistency from one student to another.

Differences in teacher assessments are broad in focus and diverse in areas. This study only focuses on examining two areas of variability and consistency. While variability indicates whether teacher assessments are lenient or stringent, consistency examines agreement in their assessment decisions. From the standpoint of variability, teachers were more tolerant with S2 while being very strict with S3. This suggests that S2 may possess some characteristics or present some features in performance that in some way incited generosity and tolerance in most of the teachers, whereas S3 may not.

In the second dimension, comparisons among teachers within students further confirms that teachers' assessments were not only different from one another, but also inconsistent across students. Despite being observed, teachers overall tended to give higher scores to S2 than the other two students; however, the difference in overall variability for each student was not extreme. Conversely, teacher differences for individual students and across students told a different story. In association with the former, the assessment performances of teachers exhibited a range of variations when the gaps between the most lenient and the least lenient assessments were quite large. These large gaps show that teachers behaved very differently and this may have resulted from teachers' individual differences. This may also be because assessments are construct-relevant (Brookhart, 2003), that is, the same performance can be perceived differently by different people. In relation to the latter, when variability levels of individual teachers across students (i.e., internal variability) were compared, three types

of behaviour were observed. The first behaviour was when one teacher assigned very lenient scores to one student but very stringent ones to the other two, introducing a possibility that their assessment may have been internally or externally biased. While internal bias refers to the interactions between the teachers and their characteristics (e.g., demographic information), external bias may be understood as teachers' interactions with variables other than themselves (e.g., students, tasks, criteria and so forth). Teacher assessment biases are further discussed in the following chapters.

In the second type of variability behaviour, one teacher's variability level remained relatively stable across all three students. This does not mean that they kept giving almost a similar score to the students, also known as the halo effect (Abikoff, Courtney, Pelham, & Koplewicz, 1993; Beckwith & Lehmann, 1975). Instead, it is because, as defined and explained earlier in this chapter, variability is the degree of difference between the mean score and the observed score and the mean scores are different for each student. Therefore, this means that the variability behaviour of that teacher was stable and, thus, predictable. Their assessments may not be influenced by student-associated variables. The last type of teacher behaviour is the reverse of the first one, in which a teacher was unusually stringent with one student but lenient with the others. Again, this may be caused externally by factors related to that student (e.g., inhibiting characteristics against that teacher's perspectives or interests). The first and the third type of observed behaviours were previously reported in literature as rater-candidate bias (He, Gou, Chien, Chen, & Chang, 2013; Kondo-Brown, 2002; Lynch & McNamara, 1998; Schaefer, 2008) that 'demonstrates the risk of the practice of single rating' (Lynch & McNamara, 1998, p. 170).

From the consistency standpoint, teacher consistency is defined as the degree of agreement that teachers reach in their assessment. Ideally, it is expected that teachers

score student performances in the same way. A student should receive a consistent score no matter how many teachers are involved in assessing their performance. Through receiving consistent scoring from different teachers, their ability in a task is faithfully reflected; therefore, the result can be relied on for fulfilling the purpose of the assessment task. However, a high degree of agreement in teacher assessments should be viewed not only as the teachers' ability to assign a consistent score but also as the process of how the score is assigned.

First, a high degree of agreement can be partly the result of teachers' sharing their expectations on how students would perform in a speaking task. Prior to assessing a student's performance, teachers are usually informed of the student's and the task's characteristics. Thus, they may presume and expect that students would demonstrate at a certain level on a range of language performance areas (Hogan, 1987; Sakyi, 2000). Moreover, it can also be that the teachers' perception and understanding of student performances are like each other. A student performance may consist of a range of components including low-level or basic components such as vocabulary knowledge, morpho-syntactic knowledge and oral text comprehension (Droop & Verhoeven, 2003) or more communicative and advanced components such as fluency, accuracy, confidence and the like. When observing these performance patterns, teachers agree with each other on how they perceive and understand those components. For example, when one teacher thinks that a student is good at maintaining the flow of the conversation by asking their partner to energise the conversation after giving their information, other teachers also observe that communication pattern from that student. This occurs across all the patterns. Finally, a high agreement in assessing student performances suggests a possibility in which teachers do not interpret and apply the assessment criteria in their own way. If the criteria are supposed to function validly and

reliably and there are no conflicts between teachers and the criteria content or structures, the way one teacher interprets and matches with student performance patterns should not differ from the others. Overall, consistency should be viewed and discussed as an interacting process of teacher presumption, perception and understanding of student performances and interpretation and application of the scoring criteria.

However, in this study, teachers were found to perform their assessment practice in several ways. The highest average consistency degree suggests that most of the teacher disagreement was caused by S1's performance and perhaps her characteristics; however, S3's performance and characteristics were observed to account for most of the agreement among teachers. Thus, teachers highly agreed with each other when they assessed the performance of S3, but they failed to agree on the performance of S1. This can be understood in several senses.

How teachers perceive and understand student performance should be taken into consideration in the very first sense. As was explained previously for differences in variability, teachers may consider their own perceptions and interpretations that are different from that explicitly described in the rubrics. For instance, some teachers may mark S1 down a bit because this student did not express an expected level of confidence due to insufficient eye contact. Meanwhile, others may assess the student's confidence as higher based on her argument and tone during the conversation.

In another sense, some classroom assessors may also consider something beyond the rubrics. For example, gender may play a role in assessment of students (Lumley & O'Sullivan, 2005; O'Loughlin, 2002). As described previously in Chapter 3, S1 is the only female student and all the teacher assessors are female. It has been reported that in cases of rater-student interactions in relation to gender, raters tend to be lenient to students of their own gender but stringent to students of the opposite gender, while other

raters do the reverse. Hence, considerable disagreement is found when comparing these two groups of raters as a whole.

In the last sense of consistency difference, the overall consistency among teachers may be attributable to the students' level of proficiency. As mentioned above, teachers tended to have a higher degree of agreement in scoring S3's performance than S1's. Interestingly, the first analysis of observed scores that is not presented in this chapter shows that S3 received the highest scores from those teachers with the mean of 3.42 out of 4.0, whereas S1 received an average score of 2.80. This means that S3 has a higher level of proficiency than S1, based on their performance in the tasks. This finding also suggests that teachers are more consistent when scoring students of higher proficiency than when scoring students of lower proficiency (Yan, 2014).

In reference to the last dimension, the difference for individual assessment categories, two main types of variations were exhibited by most of the teachers. Variability and consistency for individual categories are, respectively, the degree of discrepancy and the distance between the observed scores and the mean score for one category compared with the others. In all contexts of assessment, it is expected that the discrepancy was minimal and similar among all categories. Thus, the observed scores are not expected to be greater or smaller but closer to the mean score and the average variability and consistency for each category should be close to each other. Overall, teachers must reach absolute agreement in both variability and consistency in assessing students' performance on each performance areas. However, as reported earlier in this chapter, the variability and consistency degrees of all 12 teachers were noticeably different across categories. It has been widely reported that teachers or assessors exhibit a tendency to be stringent or consistent with one or some categories, but tolerant or inconsistent with others (Eckes, 2005, 2008, 2012; Knoch, 2011; Knoch, Read, & von

Randow, 2007; McNamara, 1996). Some assessors were unusually tougher in rating content or organisation, while others were particularly lenient in language use or mechanics (Schaefer, 2008). Therefore, the role of differences in how teachers perceived importance of assessment criteria might have accounted for their differences (Eckes, 2012).

From the perspective of variability, teachers showed highly lenient behaviours in assessing student performance on communication and strategies but reduced their extreme behaviours on the other performance areas. It seemed that the teachers may have placed more emphases on the students' ability to use the language; language-related areas such as cultural conventions, grammatical features, text structures and vocabulary may be particularly important to them when making judgements. Conversely, teachers exhibited a tendency to leniently grade student performance on communication and strategies, meaning that students had a very high chance of receiving high scores from these teachers in the language performance areas. A conclusion temporarily drawn at this point is that, overall, teachers were more lenient in assessing non-language-related performance areas than in assessing language-related performance areas. In other words, these teachers focused more on the students' ability to use the language than on their conversational ability. To some extent, most of those teachers put less weight on the two criteria in their evaluation. For these teachers, when students performed the speaking tasks it was important that they made themselves clearly understood by using appropriate strategies that they had been taught. This is not aligned but contradicts with what was reported by Cumming (2002) in his study in which ESL/EFL assessors placed significantly more weight on language use than on ideas.

From the consistency perspective, teachers' behaviour was most consistent when scoring performance on text structures, vocabulary and phonology, but more variable on grammatical features, communication and cultural conventions. The results suggest that teachers agreed more on text structures, vocabulary and phonology than on the other three categories. One of the explanations for this practice may be found in the performance descriptors in the rubrics (Wigglesworth, 1993). Contrary to what was explained for teachers' stringency on language-related performance areas, teachers' degree of consistency, as implied by Knoch (2011) and Weigle (2002) was high on some criteria but low on others and this may have been caused by the way in which the performance descriptors were constructed. For this reason, Weigle (2002) suggested guidelines be provided to develop performance descriptors for language assessment criteria. Descriptors in categories causing higher agreement among teachers may need to be more explicit and better constructed so that teachers interpret and use them similarly. Whereas, descriptors in low-agreement-causing categories constructed in an implicit manner, filled with vague ideas or a deficit of exemplars may cause confusion and conflict.

In addition, because of low-agreement-causing categories, this behaviour may be formed by teachers' comparisons between categories in the scoring criteria and the ones in their internal criteria. Teachers may assess more consistently if they had clearer and more explicit scoring rubrics that match the teachers' expectations and internal criteria. However, due to ambiguity when teachers do not agree with how an expected performance level is described, they may partly apply or totally remove those descriptors for that level in deciding on a student's ability. Instead, they switch to depending on their own criteria developed through what they think a student at a certain performance level can do and what they have done in their classroom with their own

students. A conclusion that may be drawn at this point is that, in addition to employing the assessment criteria as being asked to do so, teachers might have considered and applied their internal criteria that may not be aligned with, or may even contradict, the one they should have used. This sometimes happens when performance indicators in these assessment areas are not adequately explicit (Lumley, 2002a) and it may also happen with experienced assessors.

Furthermore, in relation to how stable a teacher's variability and consistency remained across categories, a range of variations was reported for most of the teachers. Discussions on this were initiated by Cumming (1990) and have been raised again more recently. Scoring differences across assessment categories means that a teacher exhibits unpredictable degrees of variability when scoring each of the language area. For example, teacher T01 was identified as assigning the least lenient assessment to student performances on communication and strategies, yet she demonstrated the highest level of stringency on grammatical features. Meanwhile, teachers T05, T06 and T08 displayed a different behaviour in relation to stringency, scoring text structures most stringently. Teacher T12 consistently stringently assessed performances on four out of seven categories. Some teachers were more lenient in assessing one language over others while other teachers did the reverse, resulting in scoring disagreement among teachers in the overall final scores assigned to students. These scores do not reflect the actual ability of the students. Variations within teachers may be attributable to different elements. One of the elements is in relation to teacher perceptions of the student performance areas. As discussed earlier, teachers typically place more weight on language-related criteria (Brown, 1991; Cumming, 2002; Sweedler-Brown, 1993) and this is true in the case of T01. However, T05, T06 and T08 considered text structures as the most important, while T11's emphasis was on communication, meaning that the

teachers held different views on significant language areas that contribute towards achieving a satisfactory performance even when they have a scoring rubric to follow. There might be three-dimensional interactions between teacher, student and category. This is the case when a teacher is biased towards a particular student on a particular category. For instance, T01 was very lenient with S1 on text structures and phonology while being tougher on S2 on the same language performance areas. In the same tendency, T04 was found to give the most lenient assessments to S2 on cultural conventions and strategies, while being stringent on the same criteria to S1. In this kind of bias, the teacher's emphasis on language performance areas varied from student to student. To T01, text structures and phonology may not be as important for S1 as to S2. For some reason, one or some language areas may have played the decisive role in a teacher's understanding of a student's language performance. This again indicates that teachers may have taken elements beyond language or communication performance into consideration when making their final judgement decisions.

4.6. Conclusion

In conclusion, the findings of the quantitative analysis of teacher scoring patterns show that the classroom assessors performed differently regarding variability. These findings are aligned with and are supported by several previous studies (e.g., Barkaoui, 2011; Chalhoub-Deville & Wigglesworth, 2005; Kim, 2009; Knoch & Elder, 2010; Kondo-Brown, 2002; Lumley & McNamara, 1995; Yan, 2014; Zhang & Elder, 2010). The findings also shed new light on an investigation into teacher variability in assessing EAL/D speaking works. Although variability can be considered a reality and, thus, must be faced (McNamara, 2000) and teacher assessments do not have to have complete consensus (Davison & Leung, 2009), the variability differences found in this

study should be taken into consideration. As teachers are central to the assessment process (McNamara, 2000) and are a primary assessment agent (Black & Wiliam, 1998b) they should be keenly aware of the potential biases they may have and take into consideration the variables related to students as an assessment agent (Brookhart, 2003) and assessment criteria to moderate their degree of variability. However, decisions about the real ability of students solely drawn on this source of assessment information may not be dependable. Examining variability does not only mean exploring teacher variability, but also includes an exploration into how consistently teachers carry out their roles as classroom assessors. In the next chapter, the analysis of teacher consistency will be described and explained, followed by a presentation of the main findings and discussion.

Chapter 5. Interactions in Teacher Assessment Decisions

5.1. Introduction

In Chapter 4, most of teacher participants in this study were reported to have variable assessment patterns, either internally or when compared with others, when working with new assessment material and students they had not met. In terms of variability, some of the teachers remained neutral in their assessment decisions while others were found at either ends of the continuum of variability. Regarding consistency, like variability, all 12 participants who assigned 252 sub-scores were found scattering along the continuum, indicating that there was considerable variation in assessments across participants. It was also interesting to note that among the most or least lenient and most or least consistent assessment practice, there were no correlations between variability and consistency. This means that a teacher with the most lenient assessment did not necessarily produce the most or least consistent assessment.

This chapter reports on findings from an investigation into the second research question into the factors that motivated and shaped teachers' assessment decisions. In Chapter 3, in the methodology it was explained that, to explore interactions in teacher assessments, statistical analysis would be conducted at the first instance to examine whether there were any factors that affected teachers' judgements. The factors identified from the literature for investigation were: teachers' age, current teaching position, qualifications, teaching experience and main exposure to student first languages other than English. In addition, other factors in relation to students and tasks were also examined. These are: students' gender and language background and the task genre. Following this, the findings were triangulated with qualitative data from the group

discussions and interviews that were coded both under predetermined themes and by grounded theory, to investigate for more possible factors.

The second research question involves two forms of analysis using at least four sources of data. As a first step, data were analysed using a statistical computer program called SPSS. The first source of data used in this analysis was demographic information which was collected from the questionnaire (e.g., age, teaching position, school types, main language groups, teaching experience, qualification, first language and language spoken at home). These factors were identified from the literature prior to designing the questionnaire. Another source of data used for this analysis was factors related to students (e.g., gender and language background) and tasks (e.g., task types retrieved from annotations of the three speaking tasks). As the third source of data, assessment information on teacher variability and consistency reported in the previous chapter were used as dependent factors when carrying out statistical analyses. Finally, the purpose of the teacher think-aloud protocol and interviews was bi-fold. In one way, this helped the researcher uncover new factors relating to teacher assessment decision-making that had not yet been reported in literature. Conversely, findings from this analysis allowed results from the other set of analysis to be triangulated, strengthening the answer to this research question.

5.2. Interactions in Teacher Assessments

In this part of the research study, a wide range of factors were included and examined to correlate with teachers' 252 sub-scores. The results show that while some factors were found to have effects on the assessment processes the teachers used on the day, others were reported not to have any influence on teachers' decisions. To increase logic and clarity in presenting the findings, all investigated factors were divided into

two groups including background factors and assessment factors. The factors included in the background group consists of demographic factors collected from the questionnaire (e.g., age, current teaching position, EAL/D teaching experience, TESOL qualification and main language groups). In the assessment factors group, there are student-related factors including gender and language background, language learning continuum, language components, assessment task, assessment criteria and scoring procedure.

5.2.1. Quantitative interactions with background factors

In this section, the results from statistical analyses using SPSS are first presented, indicating whether background factors had any influence on teachers' assessment decisions. Next, the degrees of influence of each background factors are reported. Finally, analyses of teachers' justifications are presented, showing a diverse source of teacher perspectives and beliefs on the effects of the background factors on teacher assessments.

It is widely agreed that assessment of productive language performance (e.g., writing and speaking), wherever and whenever human elements are included, is subject to variability (Hamp-Lyons, 1991; Huot, 1990; Janopoulos, 1993; Weigle, 2002; Williamson & Huot, 1993). Teachers as individuals are unique in their behaviours and thinking and this makes subjectivity an inherent characteristic. No matter how hard they try to be objective, neutral or bias-free when they do their jobs as language assessors, teachers may still, implicitly or explicitly, draw on a variety of considerations when evaluating a student's language ability. It has long been reported in the literature that factors that are related to teachers themselves such as experience, education background and so forth play important roles in their decision-making process, and this should

reveal itself in their assessments of student oral language development when they see and hear the actual student being assessment. In addition, this study also included in its investigation several other factors that might have interacted with such a process such as age, current teaching position and main language group. The results are mixed, showing some findings aligned with what has been reported by several other studies in the literature, but other findings were the opposite of those reported. Inclusion of new factors has also contribute some interesting new findings to the field.

5.2.1.1. Age

Although it was reported in the previous chapter that teachers significantly differed from each other when they scored student oral language skills, their assessment decisions did not seem to be affected by most of the factors reported in literature. Results from SPSS correlation analyses show that there were no significant relationships between investigated factors and teacher overall assessments in terms of variability and consistency. In terms of age, data collected from the questionnaire show that all participating teachers were over 25 years old. Specifically, five out of 12 were at least 56 years old, four out of 12 were aged between 41 and 55 years old and the remaining three were aged between 26 and 40 years old. Concerning whether teachers' ages affected their assessment, Table 5.1 shows that the Pearson Correlation Coefficients (CC) between age and overall variability is negative at $-.199$ while CC between age and overall consistency is positive at $.355$. A negative correlation means that if one factor increases the other decreases. In this case, if the teachers' age increases, their variability in assessment decreases. Thus, the older the teachers are, the less lenient or more stringent their assessments are. However, the CC is quite weak at $-.199$ and the significance level ($p=.535$) that is supposed to be $<.05$ also supports this

weak relationship between the two factors. This means that there was no effect of age on teacher variability. Similarly, a positive CC is when one factor increases and so does the other. As can be seen in Table 5.1, CC between age and overall consistency is at .355, meaning that older teachers tend to be more consistent in their scoring. However, the association is considered non-significant ($p > .05$), suggesting that the relation between the two factors happened by chance.

Although relations between the background factors and teacher assessment were considered non-significant, except for one of the main language groups and overall variability, it is important to understand how much each contributed to the variances in teacher assessments in terms of variability and consistency. To do this, multiple regression analyses were run to see the extent to which the background factors could predict dependent factors. In the first column in Table 5.2, information on background factors also includes another factor 'as a group'. The purpose of doing this is to see how much all five factors together as a group could predict variance in teacher assessment. The second column presents information about the adjusted R Square of two dependent factors. Table 5.2 illustrates that the adjusted R squares of all five independent factors as a group on overall variability and overall consistency are considerably high at .615 and .723, respectively. This means that as a group these five factors accounted for 61.5 per cent of the total variance of teacher assessment in terms of variability and 72.3 per cent in terms of consistency. Thus, they could together predict changes in teacher assessment. However, as individuals these factors did not statistically contribute to assessment variances. The adjusted R square for age is small (i.e., less than 5 per cent), meaning that this background factor cannot be used to predict teacher assessment practice.

Table 5.1

Correlations Between Background Factors and Teacher Assessments

Background Factors	Pearson Correlation	
	Overall Variability	Overall Consistency
	<i>Sig. (2-tailed)</i>	<i>Sig. (2-tailed)</i>
Age	-.199	.355
	.535	.257
Current teaching position	.313	-.065
	.323	.842
TESOL qualification	-.441	.180
	.151	.576
EAL/D experience	.504	.360
	.095	.251
Main language groups	-.674*	.387
	.016	.213

* Correlation is significant at 0.05 level (2-tailed).

In relation to interactions between five background factors and individual students and assessment categories, SPSS analyses show that while these factors interacted with assessment for one student they did not for assessment of others. Similar patterns of interactions were also found between these factors and assessment categories in terms of variability and consistency. Details of those interaction patterns are presented in Table 5.3, Table 5.4 and Table 5.5.

In Table 5.3, the first column provides information on predetermined factors and the last three columns describe details on correlations between those factors and teacher assessment on variability and consistency. Information on p-value can also be found in this table. Table 5.3 shows that teachers' age does not have any statistically significant correlations with their assessment in terms of variability and consistency, except for their assessment variability for S2. The CC between these two factors are reported to be

negative at $-.642$, meaning that the older teachers tend to be less lenient or more stringent than younger teachers in assessing oral performance by this student. In addition, the significance ($p < .05$) also proves that the relationship between age and variability for this student was real and did not happen by chance.

As significant as examining how background factors interacted with individual students, it is necessary to identify the relationships between these factors and individual assessment categories regarding variability and consistency. Table 5.4 and Table 5.5 provide information on the correlations between the background factors and the assessment categories in terms of variability and consistency, respectively. In the first column of both tables are names of seven categories including communication (V_Com or C_Com), cultural conventions (V_CulCon or C_CulCon), text structures (V_Text or C_Text), grammatical features (V_GraFea or C_GraFea), vocabulary (V_Vocab or C_Vocab), phonology (V_Phono or C_Phono) and strategies (V_Stra or C_Stra). As can be seen from Table 5.4, there are 35 interactions in total between five factors and seven assessment categories. Only four out of 35 interactions were found to be statistically significant. However, interactions between age and seven assessment categories were not found among those four significant ones. This means that this background factor did not have any effect on how lenient teacher assessment was for each of the assessment areas. A similar result is also observed in Table 5.5, in which no significant correlations between age and all assessment categories were identified.

Table 5.2

Contribution of Background Factors to Variances of Teacher Assessments

Background Factors	Adjusted R Square	
	Overall Variability	Overall Consistency
As a group	.615	.723
Age	-.056	.039
Current teaching position	.007	-.095
TESOL qualification	.114	-.064
EAL/D experience	.179	.042
Main language groups	.399	.065

5.2.1.2. Current teaching position

In relation to current employment, this study investigated whether the teachers' current working position affected their assessment practice. Demographically, four out of 12 teachers were teaching at primary schools, three out of 12 were teaching at secondary schools and the rest were EAL/D specialist teachers working with EAL/D students across both levels. Correlations between current teaching position and teacher overall variability and consistency were examined. The results reported in Table 5.1 show that there are no significant relationships between teachers' current teaching and the other two independent factors, even though two opposite COs were observed. In particular, CC for current teaching position and overall variability was positive at .313, whereas this independent factor and overall consistency had a negative CC at -.065. Moreover, not only were these COs low, but the significance value ($p > .05$) indicated that whether teachers were teaching EAL/D for primary or secondary students did not influence how they scored.

Further analyses show that this background factor did not have a significant effect on how teachers scored. Table 5.2 also shows that with the low adjusted R squares of .007 and $-.095$ current teaching position did not have any effect on teacher assessments in terms of overall variability and consistency, respectively. In relation to interactions in assessment for individual students, current teaching position has very weak correlations with teacher decisions on all three student performances in both variability and consistency. This suggests that whether teachers were lenient or stringent in assessing these students could have depended on some factors but not on the teaching position they held. Further, regarding interactions for individual assessment categories, as can be seen from Table 5.4 and Table 5.5, current teaching position did not have any significant correlations with teacher assessment variability and consistency for the assessment categories. This indicates that whether they were teaching at primary school or secondary school did not have any effect on the way in which teachers made their assessment judgements for the assessment areas.

Table 5.3
Correlations Between Background Factors and Assessment for Individual Students

Background factors	Student 1		Student 2		Student 3	
	<i>Sig. (2-tailed)</i>		<i>Sig. (2-tailed)</i>		<i>Sig. (2-tailed)</i>	
	V	C	V	C	V	C
Age	.304	.241	-.624	.372	-.212	-.261
	.337	.451	.030	.234	.506	.413
Current teaching position	.192	-.149	.041	.084	.478	-.004
	.549	.643	.899	.795	.116	.989
TESOL qualification	-.349	-.049	-.264	.199	-.238	.217
	.267	.881	.407	.534	.457	.497
EAL/D experience	.482	.160	.030	.319	.511	.111
	.112	.619	.926	.312	.089	.731
Main language group	-.294	-.157	-.603	.516	-.497	.481
	.353	.626	.038	.086	.100	.113

* Correlation is significant at 0.05 level (2-tailed).

5.2.1.3. TESOL qualification

There is an unstated hypothesis in this research that teacher assessment decisions were partly affected by their qualifications. In Australia in general and in NSW in particular, it is compulsory that ESL teachers working in public schools must have completed a four-year teaching degree or a graduate entry teaching degree or approved courses in primary or secondary teaching and TESOL. However, as mentioned earlier in Chapter 1, not all EAL/D teachers were properly trained to teach EAL/D students. Therefore, the interaction between teachers' TESOL qualifications and their assessment performance was important to investigate. In this study, most of the participating teachers (11 out of 12) were qualified TESOL teachers and one was in the process of being qualified. At the time the data were collected, this last teacher was enrolled in a TESOL degree program. As illustrated in Table 5.1, the CC between teacher

qualifications and their overall variability are noticeable but negative at $-.441$, meaning that teachers who were TESOL qualified tended to give lower scores. In terms of consistency, as described in Table 5.1, little correlation between qualifications and overall consistency was reported at $.180$. Furthermore, the connections between TESOL qualifications and variability and consistency were further examined to be non-significant ($p < .05$). This means that teachers' TESOL qualifications statistically had no effect on how they assessed student output.

The result from the regression analysis on overall variability and overall consistency (see Table 5.3) indicates that whether teachers were TESOL qualified or not did not really affect the way in which they scored student performances. This finding is aligned with that of Croninger, Rice, Rathbun, and Nishio (2007), who found no significant correlations between teachers' qualifications and their assessment of student performance. In addition, analyses for individual students in Table 5.3 show that, similar to current teaching position, the correlations between TESOL qualification and teacher assessments were low, indicating that it did not make any difference in their assessment decisions whether teachers had TESOL qualifications or did not have qualifications.

That TESOL qualifications had no statistical effect on teacher assessments could be explained as follows. As described earlier in Chapter 3, at the time participating in this study, only one teacher was being enrolled in a TESOL program while the other eleven teachers had completed their TESOL qualifications. Technically, the teacher who was enrolled may actually have done more study than those who were qualified but had completed sometime ago. For this reason, the enrolled teacher was actually qualified at the time this study was conducted. Since all teachers were TESOL qualified, their assessments were not statistically influenced by their qualifications.

Table 5.4 provides a different perspective on this background factor. Accordingly, in terms of variability, it was found that TESOL qualification had a significant effect on teacher variability in assessing student oral performance for communication at $-.784, p < .01$. However, the nature of the relationship is negative, meaning that all the teachers with TESOL qualifications tended to give more stringent assessment. In addition, this background factor was also reported to influence teacher variability when they observed cultural conventions in student work. Once again, the reported CC for this interaction was negative at $-.614, p < .05$, indicating that teachers who were TESOL qualified assessed student performance more stringently in terms of cultural conventions. However, in terms of consistency, being TESOL qualified did not actually affect the way they assessed.

5.2.1.4. EAL/D teaching experience

Although the effect of experience on assessment practice has been thoroughly investigated, debate remains among researchers and educators. Hence, this study aimed to explore the effect of these background factors on assessments by EAL/D teachers when they assessed students they did not know using assessment materials with which they were not familiar. Statistical analyses showed that teacher overall variability and consistency in teacher assessments were not driven by their teaching experience (see Table 5.1). This finding is contradictory to findings reported in several studies (Barkaoui, 2010a, 2010b, 2011; Cumming, 1990; Weigle, 1998, 1999). Conversely, it aligns with others (Leckie & Baird, 2011). However, demographic information collected from the questionnaire shows that the teachers who took part in this study were generally very experienced. For example, 50 per cent of the participating teachers at the time of data collection had been working for at least 16 years, followed by one-quarter

who had between 11 and 15 years of teaching experience. Among the last three teachers, two were the least experienced at five years and under, while the other had between six and 10 years of teaching experience. Table 5.1 also shows that the COs between teacher EAL/D experience and overall variability and consistency were positive and relatively high at .504 and .360, respectively. However, those relationships happened by chance since the significance value $p > .05$ did not suggest they were significant, implying that whether teacher assessments were lenient and consistent or not could not be predicted by their experience.

Table 5.4

Correlations Between Background Factors and Assessment Categories: Variability

Categories	Age	Current teaching position	TESOL qualification	EAL/D experience	Main language group
V_Com	-.099	.076	-.784	.562	-.558
<i>Sig.(2-tailed)</i>	.760	.813	.003	.057	.059
V_Cul	.041	.231	-.614	.512	-.422
<i>Sig.(2-tailed)</i>	.901	.469	.034	.088	.171
V_Tex	-.014	-.046	-.143	.204	-.496
<i>Sig.(2-tailed)</i>	.965	.888	.658	.525	.101
V_Gra	-.462	.465	-.199	.478	-.358
<i>Sig.(2-tailed)</i>	.131	.128	.535	.116	.253
V_Voc	-.258	.443	-.215	.345	-.716
<i>Sig.(2-tailed)</i>	.419	.150	.502	.273	.009
V_Pho	-.119	.318	-.172	.185	-.592
<i>Sig.(2-tailed)</i>	.713	.314	.594	.564	.043
V_Stra	-.125	.236	-.367	.548	-.546
<i>Sig.(2-tailed)</i>	.700	.461	.240	.065	.066

* Correlation is significant at 0.05 level (2-tailed).

Note. V: Variability, Notes: Com: Communication, Cul: Cultural Conventions, Text: Text Structures, Gram: Grammatical Features, Voc: Vocabulary, Pho: Phonology, Stra: Strategies

Like TESOL qualifications, Table 5.2 shows that the adjusted R squares of .179 and .042 for overall variability and consistency, respectively, suggest that interactions between teacher assessments and EAL/D teaching experience was not marked. However, a common tendency can be observed from the case of TESOL qualification and EAL/D teaching experience. That is, the likelihood of being able to predict teacher assessment practice based on these two factors is higher in terms of variability than it is in terms of consistency. In relation to interactions for individual student performances, as can be seen in Table 5.3 EAL/D teaching experience is the third factor to not have any significant correlations with teacher assessment for S1, S2 and S3. Thus, the number of years they had been teaching or working with EAL/D students did not shape the way in which they made their assessment decisions. Furthermore, when its relationship with assessment categories were examined, Table 5.4 and Table 5.5, EAL/D teaching experience was similar to age and current teaching position in that it had no significant correlations with any of assessment categories. This suggests differences in teacher assessment variability and consistency were not caused by the number of years they had been teaching or working with EAL/D students.

5.2.1.5. Main language group

The last background factor to be considered is the main language group the teachers were working with. There is a body of research into the effect of student language background on teacher assessments. There were seven main language groups that participating teachers were working with. Among these, Chinese had the largest

proportion being claimed by four out of 12 teachers, followed by Arabic and English each being claimed by two teachers. The other groups (e.g., Korean, Thai, Hindi and LBOTE) were the main language group for every one of the other teachers. As presented in Table 5.1, the teachers' main language groups and their overall variability had considerably strong but negative CC at $-.674$, indicating that teaching students from a certain language background did have an effect on their assessment of EAL/D oral performances. In addition, with $p < .05$ this relationship was proven to be significant, meaning that it was not randomly happening and that the teachers' overall variability level could be predicted by who they were normally teaching. Regarding consistency, a relatively strong connection was observed between the teachers' main language group and their consistency, at $.213$. However, this connection was not significant, with $p > .05$ suggesting that the effect that teacher exposure to a certain language group had on their assessment consistency with others was random. The language group that teachers were mainly working with did not decide their consistency. In terms of interactions for individual students, data from Table 5.3 reveals that, having the same tendency as age, the teachers' main language group was found to interact with their variability when they assessed S2's performance. Like the nature of correlation between age and variability for this student, these two factors were negatively correlated at $-.603$, $p < .05$. This suggests that teaching or working with students from a different language background other than English did influence teachers' assessment of the oral performance of S2, whose first language was not English.

Table 5.5

Correlations Between Background Factors and Assessment Categories: Consistency

Categories	Age	Current teaching position	TESOL qualification	EAL/D experience	Main language group
C_Com	-.013	.179	.509	.139	.152
<i>Sig.(2-tailed)</i>	.967	.577	.091	.668	.637
C_Cul	.020	-.224	.210	-.033	-.183
<i>Sig.(2-tailed)</i>	.951	.485	.512	.919	.570
C_Text	.194	.224	-.140	.110	-.329
<i>Sig.(2-tailed)</i>	.547	.483	.664	.733	.297
C_Gra	.096	.037	-.265	.468	-.273
<i>Sig.(2-tailed)</i>	.767	.910	.404	.125	.391
C_Voc	-.150	.313	-.261	.242	-.165
<i>Sig.(2-tailed)</i>	.643	.322	.412	.449	.608
CC_Ph	.323	-.549	-.077	.050	.374
<i>Sig.(2-tailed)</i>	.305	.065	.812	.876	.231
C_Stra	.346	-.343	.233	.034	.487
<i>Sig.(2-tailed)</i>	.271	.257	.466	.916	.109

Note. C: Consistency, Notes: Com: Communication, Cul: Cultural Conventions, Text: Text Structures, Gram: Grammatical Features, Voc: Vocabulary, Pho: Phonology, Stra: Strategies

When its effects on assessment categories were investigated, main language group showed a similar tendency to TESOL qualification and was also shown to affect how lenient teacher assessments were in association with vocabulary and phonology. Table 5.4 reveals that the correlation between teachers' main language group and their assessment of how students used vocabulary was remarkably high and statistically significant at -0.716 , $p < .01$. This negative relationship means that teachers who were teaching or working with students from language backgrounds other than English were

more stringent when they evaluated student performance in this particular language area. For some reason, those teachers considered this area was more important than others in defining students' ability to use the language. Additionally, teachers' main language group was also reported to contribute to their variability in assessing student phonology. Like its relationship with vocabulary, CC between this factor and phonology was found to be negative at $-.592, p < .05$, meaning that teachers with more exposure to students from non-English language backgrounds gave more importance to the role of phonology and, thus, were more stringent in their assessment in this language area.

Table 5.5 illustrates that, like the other four background factors, the main language group did not have any significant correlations with any assessment categories in terms of consistency. This means that while making decisions on student performance in different language performance areas, although the participating teachers were found to behave differently across all assessment areas in terms of consistency, their consistency was not influenced by their age, current teaching position, TESOL qualification, EAL/D experience or main language groups. This also suggests that differences in teachers' consistency across categories may have been driven by other factors that were beyond teachers' background.

5.2.2. Qualitative interactions with background factors

The purpose of this chapter is to discover what has caused differences in teachers' assessments. All background factors were investigated and were found to have modest effects on teacher assessment. On this point, the research question remains partly unanswered. To understand the entire picture of what caused the differences in teacher assessments, it is important and meaningful to give teachers an opportunity to reflect and communicate what they thought might have influenced their practice as

assessors. As mentioned earlier in Chapter 3, a group discussion was carried out after each scoring session. In addition, an interview was also conducted with individual teachers so that they could reflect on and justify their decisions.

5.2.2.1. Age

Statistically, teacher age did not have any effects on teacher assessment decisions. During group discussions, teachers were not asked to share if their age influenced their assessment or not, thus, the effect of age was only focused on in individual interviews as well as current teaching position, TESOL qualification, EAL/D experience and main language group. When asked if their age had any effect on their scoring, most of the teachers gave the same answer that their scoring could have been affected by several factors, but definitely their age was not one of them. There was a case in which one teacher was uncertain to confirm whether her age had affected her decision by explaining ‘it could have been either my age or my, you know, experience’. This result highly matches with what was found from statistical analysis that teachers’ assessment was not driven by their age. However, this finding is not aligned with previous studies confirming the effect of age on language assessment.

5.2.2.2. Teaching position

Statistical figures in

Table 5.1 show that, like teacher age, in cases in which differences among teacher assessments were found to exist whether teachers were teaching at primary levels, secondary levels or as EAL/D consultants did not really matter. Their discrepancies were not caused by the teaching position they held. However, several teachers' thought that this did affect their judgments. Although most of the participating teachers (i.e., eight out of 12) were quite certain that the teaching position they held at the time of data collection did not interact with their assessments, one-third of the participants did believe that their teaching position had an influence on their behaviours as assessors. Notably, when asked about the effect of age in assessing the students in the videos, teacher T02 observed 'I have high expectations for the students'. It is important to know that these students were in Year 10. The reason for setting high expectations for the students, she explained was 'because I've been teaching Year[s] 12 and 11 now for a long time and I've done the HSC scoring and they are overseas international'. While T02 had set high standards for these students due to her teaching at a higher level, T04 seemed to show a different angle of how her job affected her assessment practice as a research participant. T04 was a primary teacher and some of her students 'came to Australia with no English at all'. She believed that her students were still in a very early stage of learning English, and thus, her expectations were not high. She was then impressed when 'hearing students that are that fluent and can give their opinion so freely'. She noted that her teaching position manipulated her behaviour: 'So, I had to bring my opinion up a bit for all of them really'.

The other two teachers, T08 and T11, both shared that their assessment on the day of data collection was affected by what they were currently doing, but in different ways to teachers T02 and T04. If setting higher or lower expectations by T02 and T04 was the primary issue of how their teaching influenced their assessment behaviour, what

was lacking or missing in similar practice in T08 and T11's job duties was the main issue to them. Specifically, agreeing that her current job to some extent directed what she did, and it was mostly 'because I don't do that often'. Explaining her low frequency of involvement in classroom assessment practice, she said: 'I'm not full-time teaching anymore'. In a similar scenario, T11 also claimed that her job did not include assessing students in the way she was asked to do in this research. Further, she also felt that was had not focusing enough on spoken forms, explaining:

In primary school, you can actually correct students and they're quite okay about it. And you can sort of sometimes get them to maybe re-say the phrase or repeat, or something. Or you might just prompt them. But in high school it's a lot harder to do that. So, I have to admit I've been in high school for a little while now, not in a primary setting, I'm not used to, even though I might hear spoken inconsistencies, I won't often comment to the student. Because they are very self-conscious teenagers.

5.2.2.3. *TESOL Qualification*

In relation to teacher justification on the effect of their qualification, findings showed that while statistically having a TESOL qualification had no effect on teacher assessment across students it did affect their assessment practice in terms of variability on two language performance areas (e.g., communication and cultural conventions). In response, based on what teachers justified and shared, while seven out of 12 participants thought that their assessment decisions were partly affected by their qualifications, three of the last five reported no influences of qualification on assessment, leaving the other two uncertain. First, in confirmation of effects of qualification on their assessments, teachers T01, T04, T06, T09, T10, T11 and T12 strongly believed that their training from their TESOL degree programs helped them a great deal in making judgements of

student performances. For example, with a master's degree in TESOL T04 was relatively confident and explained how her qualification helped, 'I think it's given me a good starting point into knowing how language develops, how second language develops and phonological awareness and all of this sort of stuff that, I think it does help a lot actually'. Similarly, when asked how the qualification helped, T11 was absolutely positive that her degree did a good job in helping her making assessments. This teacher did a four-year bachelor's degree in primary education and said, 'Well, my primary background was all about this'. She had been trained to do assessments and at work on a daily basis she was 'objectively assessing according to outcomes, or according to criteria'. Although the other five teachers confirmed the effect of their qualifications on their assessment, they did not provide more details in their answers about how their degrees helped. However, it can be understood from what they shared that they had learned and been trained in the way in which language assessment is carried out in classroom settings.

Conversely, among the three teachers reporting no effects of their qualification on their assessment, teacher T08 denied the effect of qualification without further explanation apart from claiming 'I don't think I'm super qualified in this sort of assessment'. Conversely, the other two specifically justified their answers. For example, T02 thought that what she had been trained at university was 'not at all' useful in helping her fulfil the assessment tasks on the day. She explained that at university she was taught: 'A lot of the theory that you do at university—we created criteria, but we didn't use the criteria'. To this teacher, university equipped her with interesting theory in assessment but the skills and experience she gained at university were not very practical. Thus, she added, 'You couldn't go to a student and [say] "Okay, let's assess". You didn't do it like that unless it was on your prac [practicum]'. What she believed

really helped her on the day was what she had learned from her colleagues as she discovered, 'They're doing a lot more practical stuff now and having students involved in schools'. In another scenario, T03 completed her bachelor's degree but did not think that would have helped her on the day. According to her explanation, her qualification was not helpful in two ways. First, she thought that the qualification did not help because it had been done a very long time ago, so she might have forgotten what she had learned from it. Interestingly, her second reason was her evaluation of the degree program she took and that 'the program was not good at all'. She elaborated that the program had been primarily linguistics with a few education courses. She had not done a practicum until her very last semester, and the practicum, as she evaluated, had not been a good one. Apparently, her qualification did not help because it was not relevant with what she was doing as an EAL/D teacher and with what she did on the day. To be qualified to teach EAL/D students, she took a few short courses in ESL and discovered, '[It's] just through experience. I think the more experience and the more exposure that you get and then maybe ... more exposure to more teaching practice'.

As one of the only two teachers reporting neutral perspectives on the effect of qualification on assessment, T09, one of the most experienced teachers of this research hesitated to confirm the effect that her qualification had on her assessment. In her justification, she noted, '[I've] got an undergraduate degree with a Major in Linguistics without doing semantics, and when I look at it and I think that is so stupid, I mean language is all about meanings'. Accordingly, she had been equipped with knowledge about phonology and phonetics, so 'from that point of view, from the articulation and pronunciation point of view I kind of understand a lot of what's happening there'. However, she then concluded:

As a language specialist having a fairly sound grounding in the pronunciation, phonology side of it as well as on the grammar, sentence structure type of things, and a range of views of the grammar I feel that ... I always feel quite confident going in and talking about language.

It can be inferred from her conclusion that it could have been either what she had learned from her degree or her expertise and experience that directed her decision-making in scoring student oral performances. This comment reveals the difficulty of separating out qualifications as a discrete interactional effect in assessment, as most teachers were very experienced and qualified, and their qualifications were completed at different institutions and at different times, so the findings in relation to teacher qualifications cannot be definitive.

5.2.2.4. EAL/D teaching experience

In examining the effects of previous EAL/D experience, including teaching or working with EAL/D students, statistical analyses revealed that there were no significant correlations between this factor and assessments across students and categories. Conversely, eight out of the 12 teachers were well aware of the role their experience played in their assessments. This finding is aligned with findings reported by several language assessment researchers (Barkaoui, 2011; Brown, 1995; Lumley & McNamara, 1995; Lynch & McNamara, 1998; McNamara, 1996; Weigle, 1994; Weigle, 1998). T02, a primary school teacher who had taught English to new arrival children for more than 11 years, was acutely aware that her experience helped her in implementing assessment tasks. Similarly, with almost the same number of years of experience, T03 indicated, 'I was also using previous experience to influence where they would go'. This teacher then added, 'I think the problem is if you stay within one

setting or one context which ... for too long a time you don't get that variety and your ideas of what is a four and a three and a two and a one are skewed because they're based on your limited experience'. For this teacher, experience apparently played a crucial role in her assessment performance and; therefore, she may have mainly relied on it when doing the assessment tasks. Showing a similar tendency, teachers T05, T06, T08, T10 and T12 were positive that their experiences helped them a great deal during their scoring. For example, T05 believed that she was able to be confident on the day doing all the assessments and sharing her ideas with other teachers thanks to her 'experience using similar tools, like ESL scales and the band scales etcetera, and the learning progression'. For this teacher, if she was coming to the assessment workshop and did not know what to expect she thought it would make the assessment task much harder. Like T05, T06 thought that her assessment on that day was influenced a great deal by what she had done before assessing her own students and scoring HSC (Higher School Certificate). Conversely, T08, who started teaching EAL/D children in the 1970s, also shared that her experience to some extent affected her behaviour as an assessor. In addition, she claimed that apart from using her experience to make judgements about student performance, she also always applied her 'training in mainstream classes'. Furthermore, another strong confirmation of the effect of experience on assessment was given by T10, who was one of the two least experienced participants in this research. When asked to give her justification, she said, 'I think the experience helped because, as I said, I'm listening to students all the time and there are specific things I'm looking for, and that's based on the way that we assess using the ESL scales'. She elaborated, 'Certainly, experience, just knowledge of the cultural backgrounds of the students, and familiarity with hearing students from very little

English right through to becoming quite proficient with English. So, I think that helps me to place them somewhere’.

Finally, with more than 11 years of teaching EAL/D students, T12 agreed that her experience teaching new arrival kids and using ESL Scales helped her a lot in fulfilling the assessment tasks on that day. However, this teacher did not totally rely on her experience. Whenever feeling unsure about where she would place students, she had to make some changes. T12 remarked: ‘Yes, but sometimes I need to make some changes. I need to, yeah, this one is very good. Yeah, I need to ... depends’. Meanwhile seven out of eight teachers confirmed positive effects of experience on their assessments. T04 provided an opposite view of the raised issue. This teacher, like the other seven, indicated that what she had been doing with her students affected her assessment behaviour. However, what she experienced on the day was different. According to her justification and sharing, she did not think her experience really helped her in making judgements. When asked to elaborate on her answer, she observed, ‘in my own profession when I started, it was very much a deficit model’, implying that she has not had much opportunity to use assessment tools or work in assessment systems in which students were assessed using explicit assessment criteria. The other four teachers generally thought that their experience did not really influence what they did on the day. A typical example is the case of T11, who was the other least experienced teacher. She shared that she did not use or rely on her experience, not because she was not really experienced in doing that kind of assessment but because she compared herself with the very experienced or highly qualified teachers in the room. She explained, ‘I was extremely unconfident in the room. That’s why I think I relied more on being objective and not going out on a limb, not going out on my intuition’. She then further justified

her approach, saying, 'So, I didn't come into it with any sort of feelings of confidence. Yeah, I relied heavily on the criteria'.

5.2.2.5. Main language groups

Statistically, this factor significantly interacted with teachers' assessment in two language performance areas: vocabulary and phonology. Data from the questionnaire shows that 10 of the participating teachers had students from different language backgrounds other than English; two teachers had mainly English-speaking students as their main language groups.

Among the 10 out of 12 teachers who confirmed the effect of main language groups on their assessment, five teachers gave detailed explanations to explain their answers. For instance, given her strong exposure to students from Chinese and Korean language backgrounds, T02 confessed, 'I think maybe it did prejudice me a little bit'. From what she had exposed to, 'Chinese students are very reserved, and they just want me to give the answer ... So, it takes a long time for them to learn to be confident, like a term, to be able to have their own ideas'. Her previous exposure to students from those language backgrounds influenced her thoughts and decisions. Prior to watching the students' video, this teacher actually underestimated their ability and thought they were just like her students. After seeing them talking she realised that she had to change her view on the ability of those students because 'I was so impressed with the first girl, and the insight and the ideas that she had'. Hence, as reported in the previous chapter, she ended up by giving them very high scores. Similarly, having students primarily from Mainland China, T03 agreed that her exposure to this language group was quite helpful in her assessment on the day. Accordingly, to her the exposure helped her 'understand where common errors come from' and understanding features of their first language

helped her a great deal in evaluating their oral language ability. Likewise, T05 taught students mainly from Vietnamese and Chinese language backgrounds and indicated that she did not have any trouble understanding such students. When asked to elaborate on her answer, she noted:

When you've been teaching students from a particular language background for a long time you tune into the different intonation. So, the accent you learn to listen to, and you can distinguish what ... you can perhaps comprehend and work with their different intonations because you're tuned into it better than maybe another language that you haven't had experience with.

From her justification, the role of her primary exposure to a language group did affect the way she scored. Like T05, T10 who taught students mainly from Mongolian- and -Chinese-speaking backgrounds, claimed that her familiarity with the accent of students from those language backgrounds enabled her to understand quite well all three students' talk. She also added that her awareness of their language features helped her comprehend the students. In another example, T11, who mainly worked with students from the Philippines and Arabic backgrounds, noted, 'I think that my ability to listen is probably a little bit better than average'. Similarly, despite primarily teaching and working with English speaking students, T08 also had students from South East Asia and Arabic backgrounds. This teacher thought that she had a prejudiced view of the language ability of students from those backgrounds, commenting, 'I suppose I was very cognisant of the fact that if they're from certain South East Asian language groups like Vietnamese then their use of plurals is not very good'. She then further shared, 'I suppose that influenced me a bit'. However, the other five teachers, despite admitting the existence of interactions between their main language groups and their assessment, gave irrelevant elaborations or did not provide further explanations.

Two teachers who did not think that their assessment was affected by their main language group exposure were T01 and T06. When asked to elaborate, T01, an EAL/D specialist and consultant, replied:

I think no. I think I have a lot of exposure. You know I am aware of linguistic features, so I know how to reply. I think there's a problem and you are well aware of that as a teacher. You got to be very aware of those language features, you know you start to fill in the gaps.

She felt her ability to assess the student talk on the day was not assisted by her exposure to, or familiarity with, the accent of students, but by her expertise. In another instance, T06, who worked mainly with Chinese and Vietnamese language groups, at first indicated that she would be more sensitive to students with Chinese backgrounds. In particular, her attention was drawn to their sentence structure, expression and accent. However, she emphasised, 'It doesn't mean it would be an advantage ... because if I'm more sensitive, it's more like it will be easier for me to pick up the mistakes'. Obviously, like T05, this teacher did not deny the role of her exposure to Chinese-speaking students on her assessment, yet she did not believe her assessment was affected by this factor.

In general, in terms of overall variability and consistency, most of the background factors including age, current teaching position, EAL/D teaching experience, TESOL qualification and main language group did not have any significant correlations with teacher overall variability and consistency. However, there was an exception with significant correlations found between teachers' experience with particular language groups and their overall assessment behaviour in terms of variability. This suggests that exposure including teaching or working with students from some language backgrounds in classrooms or everyday teaching does have an

effect on teachers' assessments of EAL/D oral development. Further statistical analysis approaches found that most of the background factors did not have any significant effect on the way in which teachers scored speaking performances in this study. Although, as a group, these factors were reported to strongly influence teachers' assessment decision-making, there were weak relationships between these factors and teacher scorings.

5.3. Interactions with Assessment Factors

As mentioned earlier, one of the purposes of this research study was to discover the factors that influence teacher assessment. Apart from examining factors related to teacher backgrounds, it is also important to look for other factors related to assessment practice that could have affected teachers' process of decision-making. To do this, analysis of qualitative data was aimed to find common patterns allowing possible themes to emerge and potential factors to be identified. Thus, several factors were identified and are reported below.

5.3.1. Student-related factors

The assessment of spoken texts is reported to be the most difficult assessment task for teachers, compared with the assessment of the other three language skills: reading, listening and writing. Most of the oral assessment tasks in classroom settings are administered in live mode, that is, a teacher watches students' output and makes assessment in classes. The teacher has to pay attention not only to students' language development, but also to other characteristics she observes from watching students talking. Those characteristics (e.g., student gender, accent and personality) may distract the teacher's attention and contribute to her assessment variability. These factors will be analysed in turn.

5.3.1.1. Student gender

First, in regard to student gender, contradictory to what has been confirmed as the effect of gender in the literature (see Brown & McNamara, 2004; Carroll, 1991; Eckes, 2005; O'Loughlin, 2000; O'Loughlin, 2002; Ouazad, 2008), none of the 12 teachers believed their assessment was biased by student gender, meaning that their decisions did not depend, to any extent, on whether the students to be assessed were males or females. For example, confirming no effects of student gender on her assessment, T03 further explained, 'Because, while I felt sorry for that girl I still ... I felt like I scored it honestly according to what she could do'. And, 'It's important that when you go in you can't ... you can't have assumptions about boys being a particular way and girls being a particular way'. In another example, T09 was positive by claiming, 'I didn't know that the gender would have made that much difference, it was more the roles they chose to take ... I think that probably ... the proportion of time probably affected my judgement more than the actual gender, mainly because there was so much more you can talk about or think about if you've got more material to work with'. Other teachers even thought that students' gender might have affected their own performance when fulfilling the speaking tasks with their peers but did not necessarily affect the way they were scored.

5.3.1.2. Student accent

Second, while studies confirmed the effect of familiarity of student accent on assessment (Carey et al, 2011; Huang, 2013; Winke & Gass, 2013), in this study, only two out of 12 teachers reported that their familiarity with students' accents influenced the way in which they behaved as an assessor on the day. For example, T04 claimed of a student's accent that 'It sounded quite American. And I think that also sways me a bit

into thinking well if you have that, which is wrong perhaps'. T04 believed that on the day of data collection she was deceived by S3's impressive American accent and at the time thought that he was really good and gave him very high scores. Most of her attention was drawn towards this student's accent and, thus, his mistakes were, to some extent, not taken into consideration. Similarly, T08 also shared that students' accents influenced how she assessed students. The case of S2 was taken as an example. Accordingly, she found it hard to understand this student because of his accent even though he was quite confident and fluent in the conversation.

5.3.1.3. Student personality

The third factor related to student characteristics is student personality. Few discussions were conducted in literature regarding the effect of student personality on assessor performance. Whereas, during interviews in this study, one-third of the participants believed that the way in which they scored was driven by student personality. For example, T04 shared, 'I may have scored him a bit higher [because] ... he used humour, which I guess again comes to personality'. Agreeing with T04, T05 discovered that S2 had some of the characteristics of a good speaker because he had humour and he, to some extent, amused her. She found him 'entertaining it was more ... more of a distraction'. Another example is the case of T11. Being aware that S2 was very weak in terms of language ability especially linguistic structures and features, T11 carefully figured, 'If you don't listen carefully, his personality just makes you think oh wow he's terrific'. However, she concluded on her assessment decision that she was a bit lenient to this student 'because he's got a very engaging personality'. Likewise, T12 confessed that her assessment decision for S2 could have been affected because 'he was comical'.

5.3.2. Assessment task-related factors

Another factor that interacted with teacher decision-making in this study was the nature of the assessment tasks. As mentioned earlier in Chapter 3, three different students were engaged in three tasks of varying genres: informative, imaginative and persuasive. Teachers were asked to compare the difficulty among the tasks and decide whether the types of tasks influenced their assessment. The results show that teachers were quite clear in deciding which task type was easier to assess than the others, but did not think that, or were unsure if, their assessment was affected by that. Eight out of 12 teachers believed that the interview task, which was the persuasive task, was the easiest to score. Several reasons were given to justify this evaluation. For example, T02 noted it was easier 'Because what you're looking for is the idea of modal language'. T04 noted this was because:

The teacher knew what he was doing and, so it was a much more structured thing and I could see exactly in my mind, I knew what the task was, I knew what the, I guess the genre was and the language expectations that you have in that genre.

In another example, T09 agreed with the others that the persuasive task was simply easiest because she did not have to look at student-student interaction in the task. All she thought she had to do was look at was some very specific things she wanted to see from the student's ability to verbally persuade or 'some argument structure coming up through there'. Other teachers including T01, T05, T06, T07, T08 all believed that it was easiest to score the persuasive task. Conversely, they did not know why or were unsure about the reason.

Giving a different answer from the majority, T10 and T12 thought that the first task, which was imaginative, was easier to assess than the other two. The reason given by T10 was

There was more meta-language, so there were more things that I could use to structure my thoughts about the questioning, because the questioning was good, although it relied on the notes ... because there were a lot more of [sic] formulaic expressions.

Similarly, T12 also thought that the imaginative task was the least challenging to assess. She found it easy because, in her opinion, this task type consisted of aspects (e.g., sequencing and narrative) to which she had usually been exposed.

Unlike the other teachers, T03 and T11 did not think about which task was easier than the others. T03 believed a task was easy or difficult to score depending on the purpose of the scoring. Accordingly, if 'I was scoring for how well they've learnt the skills and the content of what I've taught, of course the informative ... because the informative is showing me how much they've understood what I've taught them'. She also thought this was, 'the same with the persuasive writing, it looks at their critical thinking'. Like T03, T11 did not comment on difficulty among tasks, but her reasoning was different. She felt that whether the task was informative, imaginative or persuasive did not make any difference. The criteria in this case played a very important role and obviously these helped her a great deal. Each task had a set of comprehensively and thoroughly developed criteria 'that gave examples of things to look out for'.

In determining the effect of task type on assessment, although most of the teachers found it more challenging to assess the informative and imaginative tasks and gave several reasons, none of the teachers thought that their assessment was affected by the nature of the tasks. To them, one task was or may have been easier than another and

they may have spent more effort to assess tasks of a certain type. However, the criteria helped to standardise their assessment even though sometimes they did not fully agree with all the performance descriptors. The task types themselves made no difference and did not affect the way in which teachers made their final decisions on student development. Given the reference to criteria, that factor was also examined to see the ways in which it affected teacher assessment decision-making.

5.3.3. Assessment criteria-related factors

First, as reported by a wide range of quantitative studies (Eckes, 2005, 2008), assessment criteria or scoring rubrics have had certain effects on assessors' decisions and behaviours. In this study, none of the participants directly confirmed the effect of criteria; however, they shared that they were influenced by factors (sub-themes) related to the criteria that they were asked to use on the day. One of the criteria-related factors was teacher evaluation of the criteria. Despite not being asked to make any evaluation against the scoring criteria, five out of 12 gave their opinion about the quality as well as the relevance of the criteria. For example, when assessing performance by S1, T03 sometimes relied on her 'gut' because: 'There were some things in the criteria I didn't like'. In this case there was a conflict between this teacher's belief about S1's actual ability and performance descriptors in the criteria. To solve this, she decided, 'like a lot of teachers we go with our gut'. In another situation, T05 discovered:

On this particular one, I actually found the criteria not as helpful as the other ones because he is obviously quite articulate and his grammatical features I thought were quite good and his text structure is quite high I thought he would come out on top.

As described earlier in Chapter 3, one set of assessment criteria was used for each task. When watching S3's performance, T05 felt that this student should have been placed at a higher level. Yet, since she was asked to use the criteria, she had to bring him down to a level in which this performance matched with the performance descriptors. Although she did not judge the relevance of the scoring rubrics explicitly, she showed her disagreement on one of the criteria. Likewise, T06 had similar responses on the criteria used to assess performance by S2. While this teacher thought, 'I just feel some of the description of the, in the scoring guidelines a little bit overlapped', she also felt that the criteria did not have enough focus on fully reflecting student ability. She said:

I think the boy is very competent in have a conversation. But, according to the scoring criteria ... the scoring criteria is more like focused on giving response. But he's very good at asking questions. I think if there is a scoring criteria [sic] about how to initiate conversation, asking questions, I give him a high score. But that one is not being assessed in the scoring.

In another example, T11 not only evaluated all the criteria, she also compared one with another. To her, the criteria used for the first task (Task 13) was quite good, even though it was not perfect, because 'it could be tweaked; it could be more obvious'. Interestingly, aligned with the evaluation by T05, T11 was not satisfied with the criteria for Task 19 and added, 'This criteria [sic] was a little bit more difficult to use, as an assessment tool'. Showing a different view on the assessment criteria, T10 was quite convinced by their clarity as well as the coverage. This teacher also believed, 'It's good to identify their next goals'.

The second criteria-related factor is teacher familiarity with the criteria. As reported in Chapter 3, training was given prior to the actual assessment sessions, so the participants familiarised themselves with criteria of analytical types. The results show

that most of the participating teachers thought that their assessment was affected by how familiar they were with the employed criteria. There were both positive and negative perspectives on how criteria familiarity influenced assessment that are then reported respectively. From the positive standpoint, T02 and T10 believed that they were quite confident in their assessment using the criteria due to their familiarity with the utilised criteria. T02 even shared her strategies to use the criteria effectively by pointing out: ‘Always look at the top, are they here, what have they got from here, and there’. Similarly, T10 explained at length how her acquaintance with the criteria facilitated her assessment. She elaborated, ‘Because I’m used to assessing using a hierarchical scale or progression, which is what I see this as—a type of learning progression or learning pathway—I think that makes it easier to understand how this is structured’. This participating teacher also claimed how the criteria used in this research aligned with the ESL Scales, which had a range of similar language area focuses. She emphasised: ‘You’re still looking along those groupings to be able to determine a student’s level of English language proficiency’.

From the negative perspective, teachers reported that they lacked confidence and their assessment was negatively influenced by the fact that they were not familiar with the employed criteria or with similar types of criteria. For instance, claiming that the criteria were new to her and she did not have enough familiarity with them, T03 affirmed, ‘I would have been more confident if I had more time with the criteria to make sure that my assessment was consistent with the larger sample’. Having a similar response, T06 reasoned that she needed more time with the rubrics so that she would be able to make reliable and consistent judgements across students. Being one of the most experienced participants, T12 was not very confident on the day because the criteria were strange to her. She commented: ‘So when I have to rate the student, I have to be

familiar with all the criteria ... I can't just judge "Oh this is it. This is where he is placed". She felt strongly that teachers must be familiar with the criteria before assessing the student. In other instances, teachers T04 and T05 both pointed out an important aspect of the criteria that affected their decisions. That is, each task had its own criteria. While they observed that this was quite good at precisely reflecting student actual ability, they also found this challenging because the matrix kept changing. They did not have enough time to become acquainted with each criterion set. By wondering: 'If the four on this one was the same as the four on the other one in a way', T04 showed her uncertainty on the validity of three different criteria she was using that was caused by her unfamiliarity with the assessment tools. Finally, despite being familiar with the ESL Scales, which were only slightly different from the criteria in this research, T09 found it hard at first and claimed, '[if] we were working straight out of ESL scales, or the EAL, the continuum, I would have been more confident because they are things that I'm used to working with'. She did not disagree with the content and the coverage of the criteria; the rubrics did not need any changes. The only thing she believed she needed to do was spend far more time examining the matrices and the samples than she and other teachers did on the day.

5.3.4. Scoring procedure –related factors

The last assessment factor reported to affect the way in which the participating teachers made their judgement decisions was the scoring procedure they had to follow to fulfil the assessment tasks. As described in Chapter 3, on the day of assessment all teachers were gathered in one room to receive comprehensive guidelines on what they would be doing. After responding to the questionnaire, they were all given a quick but thorough training session, enabling them to know what to do and to familiarise

themselves with the assessment tools and materials before they made their own assessment on three actual student performances. After finishing assessing each student performance, they were invited to share their scores with two other teachers in a group, followed by all other teachers in the room.

When asked about the level of confidence they had on the day when they fulfilled the assessment task, most of the participants responded that they were not confident in assessing student performances because of the assessment procedure they had to follow. For example, despite evaluating herself as a good listener and communicator because she had been in theatre for almost 10 years, T11 had to agree that she was not confident at all during the entire process of assessment. One of her reasons for not being confident was because of her fellow teacher-assessors taking part in group discussions conducted after each assessment section. She said, 'I didn't feel that confident, because the people in my group didn't agree with me'. Comparing her own assessment with the others would have eroded her confidence, 'so that made me doubt myself'. This must have caused a feeling of being not good at or not qualified enough to complete the assessment tasks. In addition, this teacher was also not really comfortable in the assessment setting in which she met people with high expertise in the area. She noted, 'Well, I felt very unconfident when I was in the room with [Real name] and women with PhDs who were mostly, there was women in the room older than me. Women that work with adults in the TAFE system ... I was extremely unconfident in the room'. Showing the same tendency, T08 with a very low level of confidence also explained that the assessment procedure had undermined her confidence. For instance, when asked to rate her confidence she replied, 'Not at all. I don't know the student'. She did not feel comfortable to judge tasks by students that she was not familiar with. In addition, she also added that her unfamiliarity with the assessment rubrics made her lose

her confidence. According her justification, 'The disadvantage was, I suppose, that I'm never very good at filling in criteria sheets ... I thought the criteria sheet itself was very good. It's just that I'm not very good at doing it'. Obviously, this teacher did not question the relevance, validity or content of the criteria. It is because she had not been working with it, that she was not good at using this kind of assessment matrix.

Similarly, teachers T03, T04 and T12 implicitly thought that their assessment could have been influenced by how the assessment task was conducted. For example, T03 reasoned:

I think because I wasn't so familiar with the criteria I ... like being really comfortable with the criteria before I go into the assessment. I don't think it's there for us to assess according to bands without having a lot of time with the criteria.

For this teacher, to make a trustworthy assessment she and teachers like her need a great deal more time to familiarise themselves with the criteria carefully and understand every performance descriptor in the rubrics. In another example, T04 and T12 had similar reasons for their uncertainty in their assessments. While T12, who was working mainly with primary students claimed, 'The first time I did it was my first time so ... I was not sure what to do'. Watching students talking and assessing their performance using the criteria that are different for different tasks was a new experience for this teacher. Therefore, to some extent, this new assessment mode did not give her any confidence to fulfil her job as an assessor. Interestingly, this was a new experience for T04 as well. When asked to rate her confidence she answered, 'I didn't particularly feel very confident and I think that was for two reasons mostly. I wasn't familiar with the assessment that we'd done'. T04 also added that she was not familiar with the

context; therefore, it took her quite a great deal of time to figure out the context for each task.

Indicating a slightly higher level of confidence than the aforementioned teachers, T05, T06, T09 and T10 all believed that they were affected by the assessment procedure they followed on the day. For instance, T06 thought her confidence level was just 'moderate'. Then she explained, 'One reason is because I didn't have much time to get familiar with the scoring guidelines or the scoring criteria'. Apparently, she thought she would spend more time studying the assessment instructions and comprehending all the criteria. Moreover, watching the videos, to her, for two times was not sufficient: 'I think if I'm the teacher or if I was the real scorer, I have to listen to it for two or three times to make reliable decisions'. This reason was also shared by T10, who explained, 'because we only heard them twice'. This teacher particularly had the same reason as T08 for her modest confidence when she made her assessment decisions. For her, the fact that 'I didn't know the students and their backgrounds and their context' would have interfered her decision-making process, alleviating reliability on her assessment. Another instance is the case of T09. This teacher did not think she was very confident on the day of assessment. When asked to explain, she said:

Because it was a slightly different scale than I'm used to, I found it initially a little bit hard. I mean if we were working straight out of ESL scales, or the EAL, the continuum, I would have been more confident because they are things that I'm used to working with.

Besides, like teachers with an absolute negative degree of confidence (e.g., T04 and T12), T09 also found 'reading through the descriptors here and trying to tie them into the video at the same time' challenging because she had not done similar things before. Also, like T10, for better assessment decisions, she thought:

[I] would have to sit down and probably spend a lot longer than we did looking at the samples and being ... going through probably column by column to work out where, and probably listening to, what, one, two, three, four, five, six, six to seven times to just decide where I really thought he was on.

T09 finally believed that the time that the assessment task took place could also have affected the way she scored. She noted, 'It was at the end of the session too, or towards the end, it was an afternoon session, that was it, and I was probably ... I was probably not as alert as I might have been other times'. And: 'That's a very long process and I would not often do that'.

Only one-quarter of the teachers believed their assessment practice was not affected by how the assessment was conducted on the day. Their main reason was they were quite familiar with, and relatively experienced in, assessing student oral assessments. For example, T01, one of the most experienced teachers who indicated having a high level of confidence, was quite certain that the scoring procedure did not in any way affect the way in which she scored. Accordingly, this is because what she did on the day was similar with what she normally did on a regular basis. Further, this experienced classroom assessor added, 'I think the fact that we are able to discuss with [other teachers], I really love. I think it's pretty cool to have moderation within a particular context you know within an institution'. It is obvious that the scoring procedure did not influence her assessment practice but increased her interest in doing the assessment task and, thus, contributed towards her high level of comfort and confidence. To T01, the discussion with other teachers after the administration of each assessment section was necessary and helpful. Another example of being comfortable with the way in which the assessment tasks were done is the case of T02. This teacher, having a similar reason as T01, was confident that the assessment procedure did not

have any effect on the way in which she scored. Moreover, according to her sharing, she was rather experienced and familiar with this assessment mode and assessment criteria of this type. Further, from her experience in this mode of assessment, she added:

It's better to create the expectation that the students are doing well rather than you're still down here. I think all of the kids are middle of the road in that, if I was evaluating them, I'd be really proud to be able to say they've done really well.

This suggests that teachers like her should hold positive perspectives about students' abilities prior to evaluating them. Finally, as one of the most experienced classroom assessors, T05 was certain that she was not influenced by how the assessment task was administered, except for the fact that the 'criteria kept changing'. She was used to assessment administrations in which the criteria 'strands stayed the same, so as I got more familiar with the structure I could glance at it as I was watching and sort of read to check, and I got my process better, I got better at it'. That each task had a diverse set of criteria was confusing to her because she needed to spend more time and effort understanding and interpreting the criteria. Overall, this teacher is a special case because while she was sure that she was confident and felt comfortable with the scoring administration, she implied that she would need more time to study the criteria as these were different for each task.

In general, the scoring procedure was reported by most of the participating teachers to have influenced their assessment confidence and assessment decisions. One of the most common reasons was that they were not familiar with the mode of assessment in which they had to listen or watch student performances and read the criteria and make decisions at the same time. They implicitly would need more training in doing this. Another reason was that they did not have enough time to fully understand

the students' works and the ever-changing criteria. Additionally, not knowing the students and the assessment contexts were also counted as ways in which the teachers' decisions were affected by the scoring administration. Finally, fulfilling the assessment tasks with the other teachers and moderation after assessment was also reported to affect teachers' confidence in making final decisions.

5.4. Conclusion

The purpose of this chapter was to present findings on the exploration of factors that influenced teacher assessments. Statistically, most of the predetermined factors were not found to be significantly correlated with teacher assessment in terms of variability and consistency across students and criteria. However, there was an exception in which significant relations were found between TESOL qualification and assessment variability on communication and cultural conventions and between main language group and assessment variability on vocabulary and phonology. The qualitative results revealed that most of the teachers confirmed the effect of the five factors on their assessment. When coding was conducted in a way that potential factors could be identified, several factors were identified that were related to assessment factors. Accordingly, some teachers reported that their assessment decisions were influenced by student-related factors such as gender, accent and personality. Some teachers believed that the way in which they made decisions was also affected by the criteria they used while conducting the assessment task. Finally, the way the assessment task was administered was believed by some teachers to have affected their assessment practice.

Black (2004) and Black and Wiliam (1998a, 1998b) conceptualised the classroom is a black box in which what teachers do with their teaching and student

learning is unknown. Slightly different from this perspective, the process of teacher assessment decision-making in this study can be considered a grey box, as this process has been to some extent uncovered by identifying its influential factors. However, it is important to understand not just what affected the teachers' assessment decisions, but the process they used to arrive at their final decisions and the characteristics of their decision-making style. The findings of this study relating to teacher assessment decision-making styles are presented in the following chapter.

Chapter 6. Teacher Decision-Making Processes

6.1. Introduction

The purposes of the study were to 1) examine how experienced EAL/D teachers assessed oral language development by students they did not know using unfamiliar assessment tools and 2) to identify the factors that influenced the way in which they made their assessment decisions. Some teachers indicated that they were affected by additional factors apart from the factors identified through statistical analyses. Other teachers believed their decision-making was not affected by factors highlighted in the literature but by other more task-related considerations. However, the picture of how teachers made their assessment decisions would not be complete without examining the actual process of teacher decision-making, looking for any similarities or differences between teachers in the style of decision-making Gestalt in Decision-Making Process

In classroom contexts, teacher assessment decision-making is a multi-step process by which teachers begin collecting sufficient information about student performances in classrooms and then decide where to place them on the proficiency continuum. However, as Anderson (2003) notes, ‘having good information does not guarantee a good decision will be made’. This is because the decision-making process is affected by several factors including teachers’ beliefs and values, classroom realities, external factors, teacher decision-making rationale, assessment practices and grading practices (McMillan & Nash, 2000). Due to the interplay of these factors, ‘a great amount of variety in classroom assessment and grading is evident’ (p. 31). In classroom assessment contexts, language teachers, as part of their jobs, make a wide range of specific assessment decisions about student performances on assessment tasks or

activities such as oral language assessment tasks. The decision-making process of this type is, to some extent, internal to individual teachers. In assessing student performances on these tasks, teachers first look for and gather information from different sources, for example, information about students and their outputs (Anderson, 2003). After their first reception of necessary information, their initial perceptions about student capability through their performances are formed. These perceptions are indeed essential because they inform teachers about where to place students on the proficiency continuum. Such perceptions are described in this study as assessment Gestalt—the first overall impression and perception of student proficiency and where the students are to be placed in the assessment scale. Assessment Gestalt plays a crucial role in the process when teachers make their judgement decisions. Those roles are varied in terms of importance in different assessment styles (e.g., impressionistic and analytical).

Teachers using impressionistic assessment, do not explicitly engage with assessment rubrics or rating scales. When listening to student oral outputs, teachers gain an overall impression on what students can do as a whole (Ahour & Mukundan, 2009; Wertheimer, 2012). This ‘whole’ is also referred in this study as an assessment Gestalt. This assessment Gestalt is persistently guiding teachers throughout their judgemental process. They rely on this to decide where students are in the proficiency continuum. Powered by Gestalt theory, impressionistic assessment limits assessors to accessing evidence for more particular judgements, rather than the whole (Thomas, 1994). By assigning an overall score to student performance, this Gestalt-powered assessment causes misleading interpretations of the score due to students’ ability in different language (speaking) areas (Weigle, 2002). In the analytical assessment style, assessment Gestalt acts as the first reference informing teachers of general knowledge about student outputs. This is because in analytic assessment style, teachers are required to assess

every of item of the performance or language areas and traits, instead of assigning a general score as in impressionistic assessment. To do this, teachers must rely on the assessment rubrics for individual performance areas and match these with evidence they observe from student performances. Therefore, the role of Gestalt is not fully acknowledged and, thus, remains minor in teachers' decision-making process in the analytical assessment.

Since this study is to provide more insights into the teacher decision-making process, patterns or characteristics of assessment decision-making need to be brought to light. Qualitative analyses from the interviews and discussions suggested that the ways in which individual teachers arrived at their final decisions were quite different and diverse. Based on an analysis of the data and discussions with Michell (2017), there appear to be three decision-making styles used by teachers while they assess student output. As can be seen in Figure 6.1, during the assessment decision-making process, teachers may have considered several factors that, to some extent, influenced the way in which they assessed. These factors were classified into two groups: assessment factors (i.e., foreground factors) and background factors and whether those factors influenced teachers' assessment decisions has been previously reported in this chapter. Figure 6.1 also shows that teachers with three decision-making styles started their decision-making process with their first impressions—what they were first interested in the student performances. Their first assessment Gestalt was developed based on these initial impressions. From this point, teachers with each decision-making style took different paths to arrive at their final decisions. Thus, the process of teacher decision-making consists of three stages and these are described in this section.

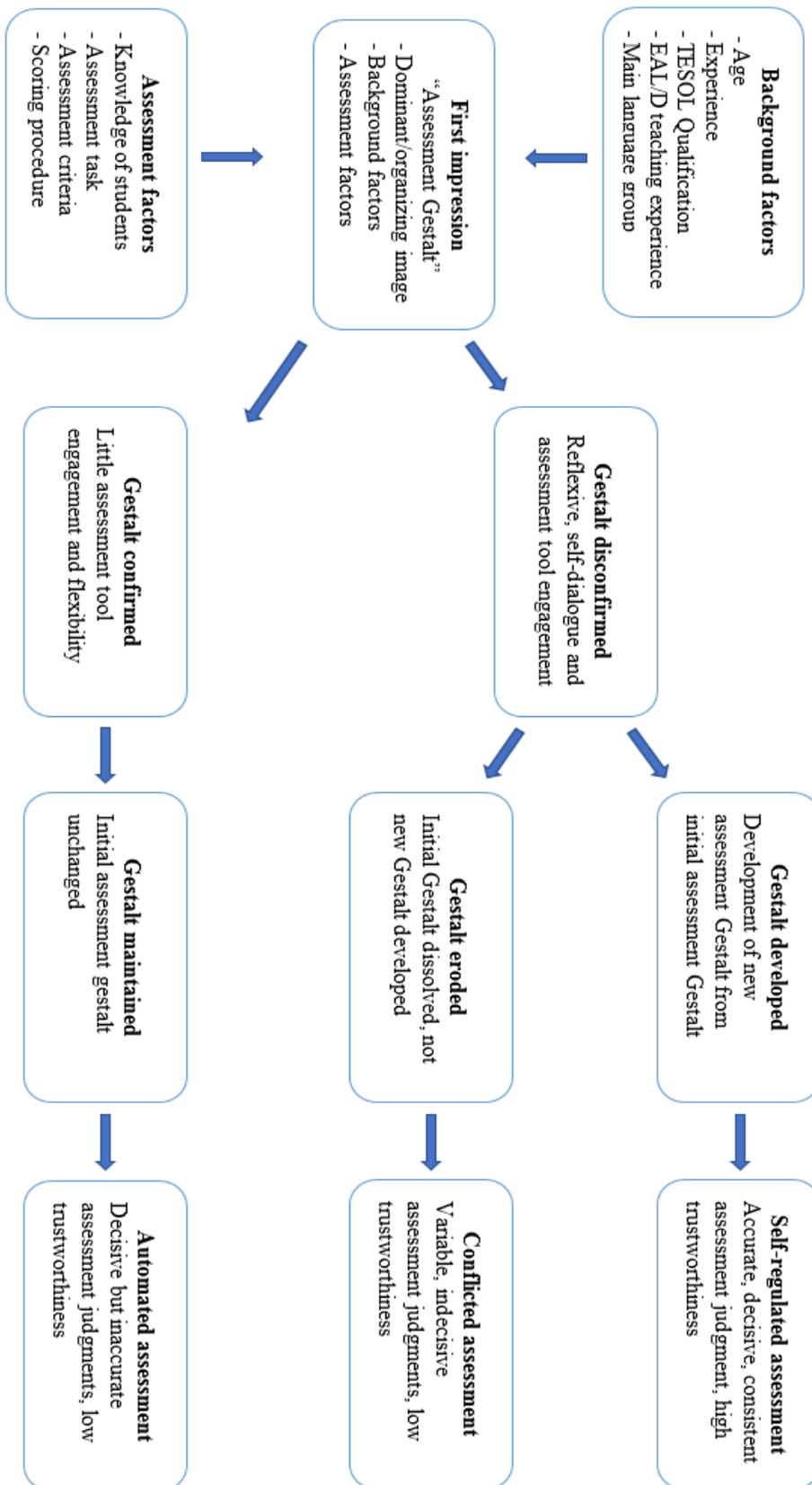


Figure 6.1 Framework of Teacher Assessment Decision-making.

In the first stage, when teachers saw the students for the first time, their attention was initially drawn to the characteristics of the student talk that were interesting to them. These characteristics could be of either the students' language abilities or the students' background knowledge. These provided the teachers with a brief overall sense of the student abilities in using the language orally. Hence, these characteristics acted as a trigger for the initial teacher Gestalt to be developed.

In the second stage of the decision-making process, after their initial Gestalt formation was triggered providing them with initial perceptions about the students' language development, teachers began to engage with the assessment tools that they were required to use. Engagement with the tools gave them opportunities to reflect and review the initial assessment Gestalt they previously developed. While engaging with the assessment tools, the teachers compared their initial perception of the students' overall performance with the assessment criteria. Typically, through this engagement process their initial assessment Gestalt was eventually changed or disconfirmed. Disconfirmation of their first Gestalt in association with assessment tool engagement made them realise that what they first saw in the student talk was not accurate; they had to change their assessment. Thus, their second assessment Gestalt about student proficiency was developed.

In this final stage of the process of decision-making, teachers had sufficient information to make their decisions about where to place students on the performance continuum. After the collapse of the first assessment Gestalt and the formation of the second Gestalt in association with assessment tool engagement, teachers had enough evidence about student abilities to allow them to make their final decision with confidence.

A sound decision-making process should include these three stages and a sound decision should not be made in the absence of any of these stages. In this study, based on teacher verbal justifications, the decision-making processes were categorised into three groups or pathways namely self-regulated assessment style with 6 teachers, conflicted assessment style with one teacher and automated assessment style with 5 teachers. Each of the groups demonstrated a different decision-making style and, in the following sections, these are presented in accordance with the three stages of the decision-making process.

6.2. Self-Regulated Assessment Style

A self-regulated assessment style in this study is defined as a process in which teachers make their assessment decisions using several sources of information but are not heavily reliant on either source. Instead, they tend to be flexible and selective in the way they use the sources. Moreover, in this study those teachers who exhibited a self-regulated decision-making style also tended to further reflect on their first perceptions and assessment tools and contrast the results with what they could observe from the reality of student performances. They used self-regulation to make final decisions. It is important to note that those teachers exhibiting the self-regulated assessment decision-making style were the only group that went through all three stages of the decision-making process. Their three stages are outlined below.

Stage 1: Trigger and formation of initial Gestalt—In the first stage, after watching the videos, teachers in decision-making indicated that they had certain impressions of student talk, including strengths or weaknesses. Those extracts of talk that stood out gave them general and firsthand knowledge of where students might be placed on the continuum. At this point, this knowledge developed into assessment

Gestalt. Their teachers generally knew where to place these students on the proficiency continuum. Figure 6.1 shows that in the first stage of the assessment decision-making process, all teachers in the three groups were similar. That is their assessment Gestalt was triggered and developed through their first impressions about student performances. However, despite having different impressions of student performances, half of the teachers were found to belong to the self-regulated decision-making style.

The first example is the case of T03, one of the most experienced EAL/D specialists. When asked to share her impressions when she first saw the students on the sample videos, this EAL/D specialist reflected that her first impression of S1 was that ‘her oral language was clunky and ... forced’. Accordingly, this was understandable because this student was conversing with two boys from different nationalities to her own. However, the teacher was also impressed with the student’s understanding of the content, noting: ‘she developed really good ideas’. This is the trigger of her Gestalt in relation to this female student. As for her impression of S2, T03 commented:

He had a really sophisticated sort of grasp of informal English. You know, he spoke confidently, he was using it really well, he wasn’t looking ... I mean he was looking for prompts in terms of making the conversation go but he was quite comfortable in responding to ... it was kind of like when he was going oh, we’ve run out of time what else do we talk about. Whereas, yeah, the girl was really clunky as opposed to [S2].

Like most of teachers with self-regulated style, T03 found that S2’s communication and interpersonal skills had a positive impact on her. This was a spark for her assessment Gestalt in relation to this student. Finally, like other teachers, her first perceptions about the last student were also very positive. The trigger of her Gestalt development in relation to this student was because, ‘He’s mastered the pronunciation,

the American pronunciation really well. So, if I saw him I'd go yeah, automatically, he's fine for entry, his oral language is fine'. For T03, this is probably because he may have been in an English-speaking environment and had used English for a long time.

Similarly, T05, who taught EAL/D to high school students and was a very experienced EAL/D specialist, had her assessment Gestalts initiated by her first impressions. However, these were disconfirmed to make room for new Gestalts formed through using the required assessment tools. Her first perceptions about the students' oral communication was not too different from other teachers. For example, for S1's talk, she noted:

Because it's also easy to be distracted by the negatives, but the detail I think, and her care, she didn't leap into it. You could see that as a negative but actually I could see that she was just thinking things through carefully before she spoke.

Although S1 displayed quite a few issues, the way in which the student took part in the conversation (e.g., starting and maintaining the conversation) impressed her. S2's communication and interpersonal skills also impressed her and made her give him a high score. In relation to S3, she rated him as quite a competent speaker: 'He is obviously quite articulate and his grammatical features I thought were quite good and his text structure is quite high. I thought he would come out on top'.

T06, as the only Chinese background EAL/D teacher among the third most experienced group, like other teachers with this decision-making style did not think she had been influenced by her first impressions of student oral communication skills. Before talking about her first impression of S1, she shared all her observations and comments about this student, for example: 'I think she was more fluent than the other boys ... because [she] used a lot of ... yeah, she used modality'. She was also really impressed with S3's overall performance and this was a sparking element for her Gestalt

to form despite the fact that the student appeared nervous. But generally, T06 viewed S3's performance as 'fluent and clear'. When talking about the S2, second student, T06 commented her first feeling about him was, 'He is very relaxed, and his intonation varied according to the conversation. Raised up his tone'. She also observed him initiating some impressive communicating and conversing strategies. However, she believed that the last student (S3) was the best. She was so impressed: 'He's very clear, fluent and both the pronunciation and the intonation are good. And [he can] give very detailed response to the hypothetical situation'. As with her thoughts on S1 and S2, she also pointed out some weak points from S3, such as 'hesitation and broken sentences', yet she believed these were minor and normal in oral communication.

Another teacher identified to belong to self-regulated decision-making style is T11, who was aged in her forties and was the least experienced EAL/D secondary school teacher. When talking about her first perception about S1, she commented:

I just liked her assertiveness but that could be ... because I just appreciated the fact that even though she is lacking a little bit of, I guess fluency with her spoken text, she really put herself out there and she butted in a bit. I liked that.

T11 also added of S1 that, 'She got the conversation going. It would have come to a stalemate or whatever; it would have stalled without her proactivity. So, I thought she was very proactive'. As for S2, she indicated that she was impressed with his very pragmatic approach and she thought this 'was a major strength for him'. She then elaborated that she felt that he was very good at engaging his counterpart and eliciting more detail from her. T11, like most of the teachers, was amazed by his communication and conversation skills. For instance, she said, 'He did it in a friendly way. He didn't do it in the same way as the first student. He was really good at keeping the conversation going, the dialogue going'.

The last teacher in this style is T12, who was the second most experienced primary teacher and mainly taught EAL/D to Hindi-speaking students. Her first impressions on the first student (S1) were that ‘She’s really trying her best to speak, although she’s not familiar with, she’s not exposed to the sounds maybe, but she’s trying to, and she uses lots of gestures to make herself understood’. As for the second student (S2), she noted: ‘The boy interacted well. But then he tried his best to explain his part there ... the way he gave ideas to the movie review’. In addition, she also recognised hesitations and limited language use from this student. Finally, she was impressed by S3’s performance noting that he was fluent and that, ‘Maybe he had been here for ... he’s not a newcomer’. She had a very good perception of the students and thought that this might have influenced her assessment of them.

Stage 2: Disconfirmation of initial Gestalt—One of the reasons for putting all the teachers in one group is that they all disconfirmed and changed their initial perceptions on student performances. The interrogation of their first Gestalt was caused by either their engagement with the assessment tools or further reflections from what they had seen. For example, when asked to decide whether her first impression had affected her assessment decision-making, EAL/D expert T03 responded promptly that she was not influenced by her first perception because all the conversations were long enough to give her time to reflect and read through the criteria while other students who were not being assessed were speaking. She then added:

The fact that it was dialogue was quite good because it forced you to also reflect back on what you had ticked and things like that. So, I think the longer piece or oral sort of sample is better and the dialogue is really good because it shows the natural sort of interactions with the other students.

If ‘the dialogue was one minute or even 30 seconds’, she may have been manipulated.

Similarly, T05, despite her positive first impression of the two students, thought she could not solely rely on this to make any decisions. To her, first impressions were good in a way that they told her what students could do but were insufficient to judge how well students were doing their jobs. Therefore, she believed and valued the role of the assessment tools and would ‘have to stick on the indicators’ in the criteria. This means her first perceptions should not be solely relied on, but, instead, used as a useful source of reference in making informed assessment decisions.

In reflecting whether her impressions had affected her assessment decisions, T06 was quite certain this was not the case. She stated:

No. Usually not ... Because I just finish assessing ... first impression may be good, because I may have ... unconsciously I have criteria set in my mind. I go, oh that’s good. But when I look at the assessment criteria, the assessing criteria I know I need to follow this standard.

This means that her first perception about the students might have been important, informing her rough ideas about students’ speaking abilities. However, there were far more important things she needed to rely on to make decisions and those were the criteria. To her ‘marking criteria are very important’ so ‘I need to know what criteria we are aligning their performance with’. If different standards are used to assess student performances, the results will vary. T06 again confirmed that even though her impression told her the students were good, she could not just provide a score. She needed to have criteria so that she could assess them accordingly.

Despite having similar impressions to the other teachers, especially the ones with self-regulated style, T11 did not think her initial perception about students’ abilities

had affected her decision-making process. Whether or not she liked the students did not significantly affect the way in which she assessed their oral works. Although there was a time when she thought that she was controlled by her thoughts on S3 when he came across ‘as more fluent and more experience[d] in the use of English’. However, she reflected:

I still relied upon the criteria. I was really pleased when I started doing it that it was quite accurate in its format. That it came up, based on my note-taking it came up at a higher level than the other students. I was quite thrilled about that and I thought this is actually quite a helpful tool.

Accordingly, her first assessment Gestalt based on her impressions changed when she saw the criteria she thought would help her to make more systematic and trustworthy assessment decisions. Finally, in her justification, T12 reflected:

But then, when I see the criteria, you know, this specifies where they are at. Because when I look at them, it’s just general. I can’t find where do I need to assess them. And then when I see this, oh this is where they should go.

What she first saw from the student only gave her general ideas about what they could or could not do, but for informed and reliable assessment, her decisions had to be made with her full engagement of the assessment tools.

Stage 3: Further reflection and self-regulation—At this final stage, due to the association of the assessment tool engagement and further reflection on student output teachers with self-regulated assessment styles developed their new Gestalt. This new Gestalt was significant for their decision-making, because it developed through careful consideration and extra reflexivity. These two provided teachers with sufficient evidence of the students’ ability to use the language so that teachers could make their final decisions with confidence.

For example, before deciding where to put S1 on the proficiency continuum, T03 carefully noted:

As teachers, when you're assessing students, you've got to be mindful of how ... because we do get fooled by students who talk the talk really confidently and things like that. Whereas the little girl [S1] her expression was not so great, but she had some really good ideas, she had some really good understanding of the text. So, I think you've got to be really careful, and if you're assessing for understanding you've got to make sure that that is weighted more and that teachers can see that.

This indicates that this teacher was well aware of the common mistakes that teachers tend to make when assessing students' oral works. Her awareness helped her to discover more evidence about the reality of the student's ability in using the language.

Similarly, in another example, T05, during self-regulating, commented that S1's strategic enthusiasm and conversation engagement may have been a factor; however, it would not have affected her general decisions. T05 noted:

It helped to inform that first communication because it was an overall judgement about the type of communication skills she had, but I don't think it affected the other aspects in terms of her strategic competence because I knew I had to look for other features.

For T05, if this student was at this level in one language area, it was important to consider if this student would be working at another level in some other skill areas. In addition, T05's self-regulation was even more obvious in deciding on S2's performance. S2's communication and interpersonal skills impressed her and made her give him a high score on this. However, she noted:

You have to step back and listen to the content and actually he didn't have a lot of content although he did have some good vocabulary, so ... but his grammatical features he had some grammatical inaccuracies which were easy to overlook because of his fluency.

Overall, during an assessment decision-making process, teachers with self-regulated decision-making style developed their first assessment Gestalt built on their first impressions triggered by some outstanding or noticeable points from the student output. Teachers with self-regulated decision-making style did not depend on their initial perceptions to make their decisions on where to place students. Their first assessment Gestalt simply acted as an auxiliary channel, providing them with first hand and general knowledge about students' ability to verbally use the language. To make consistent and trustworthy assessments, they ultimately relied on the standards, not on what they initially felt.

6.3. Conflicted Assessment Style

In this study a conflicted assessment style is a decision-making style in which teachers make their assessment about student performances based on a lack of reliable information and confidence. Like those with a self-regulated assessment decision-making style, these teachers developed their initial assessment Gestalt based on what they first saw from student performances. Following this, their assessment Gestalt changed or dissolved because of their engagement with the assessment tools and further reflection on student output. However, unlike the self-regulated style teachers who developed a new Gestalt if the first one was disconfirmed, the teacher with this decision-making style did not develop any new Gestalt if the previous one was disconfirmed. Therefore, they were not sure about their assessments even though they

engaged with the assessment tools. With the collapse of their initial Gestalt and the absence of any new Gestalt, they perceived student performances as segments, not as a whole. They also found that their belief in those performance segments conflicted with the performance descriptors in the assessment criteria. Therefore, they were reluctant to make final decisions on student development. Overall, teachers with a conflicted assessment decision-making style did not go through all three required steps of a process of decision-making in language classroom assessment.

Stage 1: Trigger and formation of initial Gestalt—At this stage, the teachers' first perceptions about the students' performances were triggered and their initial Gestalt was developed. Only one teacher, T04, was identified to belonging to this style. Demographic information shows that she was among the second most experienced primary EAL/D teachers and was among the second youngest group. When sharing her first impressions, this teacher thought that her first impressions were triggered by both students' strengths and weaknesses. Accordingly, when talking about performance by the first student, she found that this student's responses in the conversation with others were not natural, but rather were structured, formulaic and a bit stilted. Conversely, T04 thought of S1 that:

She accurately uses formulaic structures to indicate turn taking, such as what do you think. She has a broad vocabulary and is learning to use appropriate word forms. So, I think with her, part of it was that she was doing the turn taking, she had all of that, the cultural norms.

For T04, S1 was considered overall a good English user despite the fact that she did not appear very confident due to some personality or cultural reasons. The trigger for her Gestalt development of this student was her overall competence in using the language. As for her impressions for S2, she commented, 'He was definitely better than

the first one. And a lot of that had to do with the spontaneity and colloquialisms that he had'. It seems that she was impressed by S2's performance because he sounded better than the first student. Therefore, she put S2 at a higher level than S1. Similarly, by comparing student performances, her impression of the last student (S3) was that he outperformed the second student: 'He was self-correcting as well which was very good. They all did a bit. And it also helped that he's developed a bit of an accent as well that is a native like [sic] accent. It sounded quite American'.

Stage 2: Disconfirmation and dissolution of initial Gestalt—In the previous stage, the teacher (T04) had some ideas about where to place the students in the continuum. However, in this stage, this teacher's initial Gestalt collapsed when she became more engaged with the assessment tools. For example, she said, 'Like I said before, for instance, that last student, well the second student, he was just so funny, and because he's so confident ... then the criteria grounds you'. Reviewing the criteria did make her examine her first Gestalt and activated her flexibility. She noted, 'You start looking at, what about their verb endings, are they using modal verbs, are they just using formulaic language. I think that is very important to come down'. She also reflected on her experience from her own work: 'Someone has very good vocabulary, but their phonology is still behind. And I think that was the case with at least one or two of these students as well'. This indicates that the criteria and her experience told her that what she first observed could be a reliable source of reference but could not be solely depended on.

As previously mentioned, the teacher with this style did not go through all three stages in her decision-making process; therefore, her decisions were made with the absence of Stage 3. When asked to decide whether her first impression influenced her assessment decision-making, T04 responded with conflicts in her own answers.

Notably, she said, ‘Yes, well, quite a bit I think’, meaning that she was somewhat aware of the effect of her first impressions on her decisions. The effect was indicated in that ‘If I had to give the students a one to four, they’d all probably be a bit higher’. In cases in which student performances lay in the borderline between two bands, her impressions would make her go for the higher band. Furthermore, this teacher she did think there may have been interactions between her first impressions and her assessments. She observed: ‘I think, as a reflective teacher, that I would have to be a bit dishonest to say that I do not have biases. And maybe they’re not conscious, but I think everybody does’.

Stage 3: Confused decisions—Once T04’s initial perceptions about student performances were disconfirmed, she became confused about aligning what she saw from student performances with what she understood from the criteria. In her justification, she seemed indecisive. For example, when she had to assess S1’s performance, she decided that this student was halfway between a two and three: ‘If I can’t decide I should always assess them down’. This is contrary to what she had said earlier when she indicated that she would give higher scores for students on the borderline. Thus, in the end, ‘That’s how I reached that decision ... I went “Okay, she’s halfway in-between so I’ll go for two”’. In another instance, when assessing S2’s performance, although T04 found he was very confident and she wanted to give him a four, ‘in the end I felt that I couldn’t, based on the criteria’. This means that this student had met the criteria at level four, but she had not put him at that level because she felt he might not be that proficient, compared to the last one. So, she had not fully engaged with the assessment criteria. Regarding the last student, in deciding on his performance on one of the language areas, she was again uncertain which way she should go. She observed:

I couldn't decide ... I gave him two and then I changed it back to a three and I couldn't really decide for that one. And that probably dragged him down a little bit as well. I think if I'd been confident that that was a level three, then maybe I could have pushed him up a bit more.

She made her final decisions in the absence of stable reliable information.

To summarise, with the collapse of the first assessment Gestalt and the absence of new Gestalt, the teacher with conflicted decision-making style made decisions based on her vague and uncertain perceptions. Decisions made that rely on vagueness and uncertainty may not be dependable, resulting in low trustworthiness.

6.4. Automated Assessment Style

In an automated assessment style, teachers watched student performances and developed their assessment Gestalt through their first impressions of them. However, these Gestalts remained unchanged throughout the entire assessment process regardless of the presence of the criteria that were supposed to be the benchmarks for the final decisions. This meant that during the assessment process, teachers with this decision-making style did not have or had minimal engagement with the assessment tools, instead they held onto their initial perception and, thus, their decisions were mainly based on their first perceptions which tended to be formed by experience (Barkaoui, 2010b). Hence, teachers with this assessment decision-making style experienced all the stages of a decision-making process; however, there were no differences between the last two stages and the first stage. This is because in the last two stages their initial Gestalt remained dominant and shaped their final judgements. In this study this kind of assessment practice is called 'automated assessment' and normally results in decisive but inaccurate assessment judgements. Therefore, such assessment is considered low in

terms of trustworthiness. Interestingly, more than one-third of the participants were identified to have this automated decision-making style.

Stage 1: Trigger and formation of initial Gestalt—Teachers with automated decision-making style provided various perspectives regarding the ways in which their impressions were triggered, and their initial Gestalt developed. The first example is the case of T01. As the most experienced EAL/D teacher and specialist, T01, during the interview, saw herself as an expert in the field although her students were at a higher level than the students in this study. When asked about her impressions of the student oral communication skills, she commented that she had appreciated S1 because ‘At first she was very confident. She presented a very diligent student who’d really gone over the material. She’s obviously familiar with that. Her articulation, you know she opened her mouth and articulated’. Additionally, she was really impressed that S2 ‘was a very skilled communicator. And very engaging and, you know, he’s got a lot of personality, very interested in people. He was very observant, he’s watching the person he’s communicating with and reading memos’. Her impression of S3 was general, in that she found this student was ‘particularly good’ and his strategy to use modality was ‘to some extent, but very impressive, very confident’.

When the question of whether her first impression of student performances influenced her final decision for those students was raised with her, T01 shared that, as a classroom teacher, she was used to probing student works. Yet, when she started to see through different criteria for all the tasks, although their ‘paralinguistic features can be very persuasive’, she emphasised that ‘My job is to work beyond that. I like to work beyond my first impression’. Based on her initial justification, her first impression acted as a reference channel to form her first perception of student proficiency and the criteria helped her come to stabilised judgement decisions. However, she concluded her first

impressions of students had ‘strong influence’ on the way she arrived at her overall assessment decisions. Overall, it can be concluded from T01’s comments that, ideally, she wanted to work beyond her first impressions, but during the administration of the assessment her practice was dominated by her initial impression of the student’s oral output.

Another example is T02, who was an experienced secondary school EAL/D teacher who taught mainly Korean-speaking students. When she was asked about the first interesting thing she saw in S1’s performance, she noted:

With the girl, I was impressed at how she did throw a bit of insight into the ideas of the film. It wasn’t just a black and white ... she was able to counteract. I thought that was really good. She was clever, I thought.

Obviously, the way in which this student interacted and argued drew her attention first. Similarly, regarding S2’s performance she thought: ‘One of his strengths was in the way he spoke. He did sound colloquial, but because it wasn’t too formal, and I think that’s how your attention [was] a bit with his conversation, [not] with the girl’. Finally, her impression of the Mongolian background student, S3, with the North American English accent was ‘his pronunciation of words’, although she found ‘he was a little boring in his responses’.

Responding to the question of whether her assessment practice was affected by her first impressions, this experienced high school teacher commented, ‘I know you’re not meant to compare students. You’re not meant to compare, but it is really hard not to’. For this teacher, holding her first impressions and then comparing the performance of students she was impressed with was what primarily influenced her assessment decisions. For example, regarding S1’s performance, she noted:

When you look at the first group, the three students sitting there together, one thing I did like [was] how the girl held the conversation ... So, I think that would influence me in terms—even though I know we're probably meant to assess language skills, but I think she was very good, and that's why I would be more influenced for her.

Similarly, when comparing S2's performance with the performance of the girl (S1) in the same task, T02's first impression made her sympathise with S2's inferior performance. For instance, she found:

The girl had good answers. She knew what she was talking about. She had a lot of knowledge about the characters. More so than what he had ... but he displayed more confidence in the way he was speaking than the girl. She sat quite still, whereas he was leaning all over, which I think is a street, smart kind of kid. He didn't have the formality in the same way as the girl did, but that could be part of his personality as well, because people have different kinds of personalities.

In terms of the effect of her first impression of S3's job interview, she commented that he did not interact with any students and her initial impression of him was his boring responses. Consequently, she focused on his drawbacks when assessing his performance. To illustrate, she reported:

He answered the questions, he did what was required of him in terms of applying for that position. He could have done a bit more, definitely ... adding to his own ideas and his own personal experience. When he was talking about he's done, some acting in his home country, he didn't really say what it was that he did. Was he the main actor in the play or the singing? He could have added a bit more to that, which he didn't do.

T02's assessment of this student, as reported in Chapter 4, was the second lowest compared with the assessments of other teachers for this student. This means that her assessment Gestalt for this student from her first impressions of him remained unchanged and was a dominant factor in her decision-making.

The third teacher identified with this style, T08, is like T01 in that she is an experienced EAL/D specialist and consultant. Responding to the question about her first impression, T08 commented that, from the moment she watched the video of the first student for the first time, she found, 'She clearly knows how to interact in a discussion. So, her strengths are that she knows what an oral discussion is all about'. However, T08 felt that S1 would need a great deal more vocabulary to be at a higher proficiency level. As for the second student's performance, T08 found that S2 had 'an engaging personality in an oral discussion'. T08 felt that what S2 really needed was vocabulary to be 'a very articulate, engaging speaker'. This was similar to what she saw in the first student's performance. Again, when sharing her first impression of S3, T08 noted:

He's confident. He appropriately avoids negotiating and communicating. I think it's quite clever. I'd do the same thing. I think he does it well. So, I don't think it's a defect. Look he's able to have a very sustained conversation.

Contrary to what impressed T02 about this student's performance, T08 observed that S3 was quite successful in his role in the interview task. She observed that, 'He's prepared to take a risk. He knows what he's meant to do. He clearly engaged with the person. And so, he was able to do a lot of the communication cultural conventions and all of that'.

When reflecting whether her first impressions affected her assessment decisions, T08 thought, 'That could very well happen'. It is also interesting to observe in her explanation that her first impression helped her a great deal by giving her a rough idea

of what level the students were at. However, she made changes to her ideas, meaning that her first impressions could be changed. Interestingly, her first impressions, to some extent, influenced the changes. Thus, her first impressions helped in forming her very first understanding of student proficiency. Then, when she listened to the performances for the second time and read through the criteria, she decided to make some changes to her initial understanding. Changes were made mainly based on her very first impressions. Therefore, it can be concluded at this point that the entire decision-making process of T08 would have been shaped by her first perceptions about the students' works.

Sharing commonalities with most of the teachers with self-regulated style, T09, the most experienced EAL/D primary teacher in the highest age range, found herself influenced by her first impressions of student performances. When sharing her first impression of S1, she noted: 'She was the type of student who would take a leadership role in any group work'. To some extent, her impression of this student was similar to the other three teachers. They were all especially interested in the way in which S1 took charge and led the conversation. The fact that 'she seemed to take charge and seemed to be very competent', led T09 to believe that she might have scored S1 higher. As described in Chapter 3, prior to the oral justification and interview, teachers had a chance to watch the videos a second time to provide them again with information on student performance so that the teachers could better reflect on their assessment practice. Listening to the student again, T09 noticed that she had not realised or had ignored grammatical issues in S1's performance on the day 'because she was providing so much information and doing it reasonably articulately'.

In terms of her impression of S2's performance and its effect on her decision, sT09 e commented that although S2 had demonstrated limited talk time and several speech problems, she found:

He had a whole lot of the non-verbal[s] and his ... he was the perfect talk show host. ... and he had a lot of the ... even the gestures and the ... and the demeanour of a talk show host in talking into an interview ... into an interview guest.

To T09, S2's communication and conversation abilities were good and noticeable. Therefore, 'It would have influenced me, then'. During her reflection, she found that she had also seen several weak points in S2's talk; however, she had not marked him down because of these. Instead, she had given him 'a relatively high score' and explained, 'I might have been feeling very generous that afternoon'. This means that T09's assessment decision was affected. Despite being aware of the drawbacks in the student's performance, she assigned her final score to this student mainly based on what she was first impressed with. Finally, with the last student, S3, she was quite impressed and convinced by his near-native American accent. Although during the interview she did not respond to the question about the effect of her impression on her assessment, she spent most of the interview time talking about what she was impressed by with this student. She did not recognise much trouble in his speaking; therefore, she put him mostly at band four across the performance areas.

Stage 2 and 3: Persistence and dominance of initial Gestalt—After their initial Gestalts were developed from their first perceptions of student performances, these teachers' seemingly automated Gestalts remained stable and determined their final decisions. Unlike teachers with self-regulated and conflicted decision-making styles, their initial understanding of what students could do in the first stage of assessment

prevented or limited those teachers from collecting more information to facilitate a better and more precise understanding of the students' talk. There are several possible explanations for why these teachers did this. The first explanation is that these teachers relied heavily on their initial impressions and were confident that this understanding of students' abilities was correct and could be depended on in decision-making. This suggests that these teachers considered it unnecessary to obtain more information about the student's language development because they were quite confident with what they saw for the first time. Thus, engagement with the assessment tools was almost redundant. The other possibility is that these teachers did look for more evidence about students' abilities, but their initial Gestalts resisted what they found in the assessment tools. Their confidence with their initial Gestalt may have blocked them from seeing the different perspectives that the assessment tools provides about students' abilities. Besides, their initial Gestalts might have made these teachers reject part or all of the information provided by the tools if it contradicted what they had initially observed.

Overall, in decision-making style, teachers' first perception about the students' performances play a decisive role in making their final decisions. This assessment decision-making style is very similar to the impressionistic or holistic assessment style that has received a great deal of research attention (Carr, 2000; Mitchell, 1996; Tyndall & Kenyon, 1996; Vaughan, 1991). During their assessment process, those teachers may have discounted the assessment rubrics that were supposed to be the main assessment tools, and instead relied on their initial perceptions. As noted by Anderson (2003), perception of student performance is not the reality of student ability, but is simply a lens to see such reality. Teachers with automated assessment decision-making style relied on what they see through a single lens to decide on students' ability. As McMillan and Nash (2000) suggest, decisions that are made with a lack of reliable information are

usually not dependable. Therefore, these teachers' process of making assessments as well as their judgments may be somewhat questionable in terms of trustworthiness.

6.5. Comparisons of Teacher Assessment according to Assessment Styles

This study has demonstrated that teachers from different groups demonstrated different assessment styles when making their decisions. This raises the question of whether differences in teacher decision-making styles were influenced by background factors. Therefore, further analyses and comparisons among the three groups are necessary. In this section, comparisons among teachers with three different styles are presented in four different aspects such as demographic differences, distribution of interactions with factors, disagreements in perception and differences in terms of variability and consistency.

6.5.1. Demographic differences among assessment styles

Demographically, teachers with all three decision-making styles were different from each other in age, experience and teaching position. First, teachers with the self-regulated style had the highest mean age range at 56-above, leaving teachers with conflicted and automated decision-making styles at 41-55 and 26-40 respectively. Similarly, the self-regulated style teachers were again reported to be more experienced in teaching or working with EAL/D students than the other two groups. The mean experience range for this group of teachers was 16 years and above, followed by the other two groups in the same order as age. Finally, two out of five teachers with self-regulated decision-making style were EAL/D specialists working with EAL/D students, a similar number of teachers were teaching at primary level and one was a secondary

school teacher. As for the automated style group, while half of the teachers were EAL/D specialists, two were teaching secondary school students and one was a primary teacher. The only teacher with conflicted decision-making style was teaching EAL/D for primary students. Thus, in terms of teaching position, on average the teachers with self-regulated decision-making style were teaching more mature students than the teachers with automated decision-making style. The conflicted decision-making style teacher was teaching the youngest students of all. Teachers with three decision-making styles demonstrated different approaches in reaching their decisions, so there may have been an influence from their backgrounds in their decision-making.

6.5.2. Distribution of interactions with background and assessment-related factors

Teachers with three different decision-making styles were found to be different in distribution of are shown in Table 6.1. In the first two columns in Table 6.1 below are teacher groups according to the three assessment decision-making styles. All factors investigated are presented in the third column. All 11 factors are labelled using their initials. For example, A stands for age, CT for current teaching position, Q for qualification, E for experience, ML for main language group, SG for student gender, SA for student accent, SP for student personality, T for task, C for criteria and AP for assessment procedure. These interactions between teacher assessments and other factors were derived from the teachers' justifications and sharing in group discussions.

Table 6.1

Differences among Decision-making Styles: Interactions with Influential Factors

Styles	Teachers	Factors										
		A	CT	Q	E	ML	SG	SA	SP	T	C	AP
Self-regulated	T03					x					x	x
	T05				x	x			x	x	x	
	T06				x	x				x	x	x
	T07				x	x				x		x
	T11		x	x		x	x		x			x
	T12				x		x		x	x		x
Conflicted	T04		x	x	x	x	x	x	x	x		x
Automated	T01			x						x		
	T02		x		x	x		x		x	x	x
	T08		x	x	x	x				x	x	x
	T09			x	x	x				x	x	x
	T10					x				x		

Note. A: age, CT: current teaching position, Q: qualification, E: experience, ML: main language group, SB: student gender, SA: student accent, SP: student personality, T: task, C: criteria, AP: assessment procedure.

According to Table 6.1, assessments by teachers with automated decision-making styles seem to be influenced by almost all factors, except for age, student gender and student personality. Specifically, there are 24 interactions across all factors observed for the group of teachers. The number that is a little higher is that for teachers with self-regulated decision-making styles. This group has 28 interactions with almost all factors except for age and student accent. In addition, it is noticeable that the only teacher with a conflicted decision-making style has the most interactions with almost all groups of factors, compared to the other groups of teachers. Her assessment decisions were affected by nine out of 11 reported factors. Therefore, differences among teachers in

terms of their decision-making styles may have been the consequence of how much they were influenced by teacher background as well as by the assessment itself.

It is also significant to observe patterns for each group of teachers in the distribution of their assessment interactions. Specifically, the automated style teachers tend to interact equally with both background factors and assessment-related factors, meaning that when they are asked to carry out similar assessment tasks, both kinds of factors are more likely to have the same effects on their assessment practice. No group of factors outperforms or has a stronger influence on teacher decisions than another. Conversely, both the self-regulated decision-making style teachers and the conflicted decision-making style teacher are similar in that they all tended to be influenced more by assessment-related factors than by background-related factors. This means that their assessment decisions are more predictable and explainable by factors related to students, assessment tasks and assessment administration. These patterns of distribution of interactions among the three groups of teachers may help somewhat in explaining the way in which they made their assessment decisions.

6.5.3. Disagreement in perceptions among teachers

As mentioned earlier, differences among teacher assessments may have resulted from teachers' individual differences. One difference is whether they agreed or disagreed with each other in perceiving student oral outputs. For this study, only disagreement among teachers was investigated and the data were collected from teacher discussions. It is expected that teachers would agree with each other not only in their final decisions, but also in the way in which they perceive and understood students' output. The results from the group discussions indicate that in most cases this was true, in that teachers came to an agreement on how they understood individual student

output. However, in several instances teachers were observed to disagree with one another.

Most of the disagreement among teachers related to their opinions about students' strengths and weaknesses. While some teachers considered an area of student's talk a weakness and were willing to mark them down, others argued that the same area was normal or even a strong point worth crediting. Before giving examples of teacher disagreement, it is important to note that the teachers were grouped randomly for group discussions after each assessment session. For example, teachers T07, T08 and T10 worked in the first group; T01, T05 and T06 formed the second group; T09, T11 and T12 formed the third group and the fourth group consisted of T02, T03 and T04. For the study, disagreement among teachers were categorised into two groups: productive disagreements and unproductive disagreements.

Productive disagreements: productive disagreements are referred to as cases when, during their discussion with others, teachers disagreed with each other in their perceptions of student talk. Throughout the moderation process, teachers were open to ideas from one or two other teachers in the same discussion group. They were willing to come to an agreement if they found others' arguments were convincing. One example of productive disagreements was the case of T03. In discussion with T02 and T04, T03 at first strongly disagreed with them on S2's use of gestures in his conversation with the female student. An excerpt of their discussion is provided below.

T02: Sometimes, you know, he did use the gestures and I ticked at two, he did that so well, gestures, showing his fingers out when making a point.

T03: He didn't use gestures as a strategy I felt. I just thought he used his natural body language, you know what I mean?

T04: I thought it was a flow any way when he used the gestures. I guess he was confident in using the language by using hand gestures. It came natural to him and he was still able to use them because he's sitting down you know, using hand gestures is difficult. It's more natural to do that.

According to T02, using gestures during the conversation was a strong point of S2's and she gave him credit for this. In the same way, T04 believed that using hand gestures was part of being an effective communicator and she also noted that sitting and using hand gestures was not easy and this strategy was one of his strengths. However, T03 did not agree. She believed using hand gestures was his body language and that this student used them as a habit in conversing with others. Therefore, she did not and would not give credit for this. T03, after hearing T02 and T04 justifying their perceptions about S2's hand gestures as an effective communication strategy instead of one of his personal habits, agreed to make changes to what she had initially thought.

T03: I think I am just getting defensive because I think that because I thought he's a two, but he's not a two (laughing).

T02: Yeah. He's a three (laughing).

T03: Yeah (laughing).

T04: I go to a four with him there. That was with the beginning. I really liked how he had some of those strategies.

In another example of productive disagreements, working in a group with T01, T05 and T06 in most cases agreed with each other. However, their opinions on how S2 was involved in the conversation were in conflict. An excerpt of their disagreement is featured below.

T06: I mean he's just got one role, that's peppering her the questions and she's doing all the heavy lifting and asking the questions. But there's a kind of a symmetry, he's ... You're not really seeing him do the full ...

T01: But, it was a strategy.

T05: But, I'm wondering though whether that was actually what the task was or whether the task was – it always comes back to the purpose. And I think you need to see the purpose, because you need to see whether he was given the same role as her.

T06: That's how it worked out. So, he might have taken the strategic role to avoid the heavy lifting and ...

T05: Maybe. But, I also thought maybe what he was doing was using the questions because it was part of being an effective communicator.

According to the above excerpt, T06 believed S2 did not perform well in conversation with the girl, because most of the time he tried to avoid the difficult part of the conversation, demonstrating ineffective communication. However, T05 and T01 did not agree with T06 on this point. T05 did not classify this as one of the student's weak points and thought that he did not avoid 'hard' work in the task. She believed that asking questions was S2's attempt to be an effective communicator. On this point, T06 was convinced and adapted her opinions to agree with T05 and T01, she noted, 'part of strategic confidence is managing to avoid the discourse that you're not very good at'. At first, she ascertained that S2 did not want to take the difficult role, but after moderation she agreed with the two other teachers in her group that this was part of his strategic competence.

In another example, T07 disagreed with T08 about S2's involvement in his conversation with the other student. Below is an excerpt from their discussion.

T07: So, I thought there was a point where he did rely quite heavily on it but then he went back to, that I think I was making fours, some of the interaction. When they were having that more natural interaction he wasn't referring to his notes.

T10: But, he had quite a lot of questions to ask her, didn't he?

T07: I mean I wouldn't remember.

T10: I wouldn't have remembered all the questions either. I would have been going back to look at the notes.

T08: Same here.

T07: So, what level did you give him?

T08: Well, I just gave him a one but now I'd like to change it. But I think because I ...

T07: You're tough.

T08: No. Because I think the pedagogy of the teacher will influence how a child doesn't assess them.

T07: But aren't we just assessing the ...

T08: No, I know. I know that, but I thought that at first, he thought he just had to look at his notes. And then when relaxed he just deviated from his notes.

T10: Really? It was quite hard, I thought, because he was just asking her questions all the time. There wasn't really enough.

As can be seen in this excerpt, teachers T07 agreed with T10 who believed that S2 mainly relied on his 'notes' when the conversation first started. When it became more natural and flowing and he felt more relaxed, he gradually deviated from them. However, she disagreed with T08, who held an opposite perception towards the issue and commented that she was not convinced that this student was actively involved in

the talk, always using the same phrases or sentences to converse with the girl.

Eventually, however, T07 decided what T08 felt was reasonable. So, she thought she would change her decision, demonstrated in the excerpt below.

T07: But there was a point that ...

T08: That's what I mean, at first and then suddenly they started engaging with each other and ...

T10: Then they started talking. I mean he was really reading what was on the ...
What did you think of the movie? Did you like the end?

T07: I mean he did tend to engage her in further conversation about her ideas but because he had the same sentences. Is that what you ... I don't know if I thought of it as ... is that what you really think it was something like that. But it was kind of the same phrase that he had to try and prompt her for more information.
So, you're correct, and that makes me re-evaluate mine.

As also seen in the above excerpt, T10 indicated she benefited from the discussion with the other two teachers. Like T07, T10 at first did not really agree with T08 about S2's involvement in the conversation with the girl. She found that apart from asking questions, the male student (S2) did not do enough in the conversation to show his language ability. However, on hearing T07 identify a convincing point by T08, T10 started to agree with T08's argument. She noted, 'Then they started talking', meaning that the two students did really talk to each other, they were not just asking and answering.

The last example of productive disagreements is the case of T11 and T12 who argued with each other over S2's speech issue. Below is an excerpt from their discussion.

T09: He possibly has some kind of physical issues. I think [it] might have to do with the tongue. He might have a longue-tie.

T11: I mean other than that I think his pronunciation has been a bit clearer.

T12: That's cultural, because I heard many Chinese students speak that way unless they were trained to speak more clearly. Especially those who can speak in English, they can't speak in Chinese.

T11: Really? I have many Chinese students and I don't think they would need learning support.

T12 thought that these issues were typical of Chinese-speaking students who would need a great deal of learning support. Again, T11 disagreed with this and thought that these issues were normal for students from language backgrounds other than English and claimed, 'I do not think they would need learning support'. However, in the end T11 found that T12 and T09 had a point thinking S2 may have had a speech issue. She said, 'I think he may have a speech impediment. I am not sure'. She gave ground, apparently convinced by T12's argument.

6.5.3.1. Unproductive disagreements

Unproductive disagreements are cases in which teachers, rather than just disagreeing with each other on aspects of the student talk, and their disagreements remained unchanged after sharing and discussion, do not respect or listen to each other and refuse to engage with each other to justify their assessment decisions. Hence, teachers may not have really benefited from the moderation. The first example of unproductive disagreements is the case of T08. As can be seen from the two excerpts from her discussion with T07 and T10, in the entire time T08 defended her opinions over the issue that S2 actively engaged in the conversation with the first in the movie

review task. She thought her perception about S2's involvement was accurate and tried to convince the others to agree with her. By doing so, she may have prevented herself from seeing the good points from T07's and T10's contribution to the moderation. This suggests that moderation did not really help with her decision in this regard.

Another example of unproductive disagreements was T11. While T09 seemed to agree with T12 on her perception and decision and did not really comment on S1's use of tense, T11 explicitly indicated her disagreement with T12 over the issue. The excerpt of their disagreement is below.

T12: She expressed suggestions, agreement and she had eye contact with partners, but sometimes she looked away for some reason. So, I put her in level two. The turn taking was moderate. She used present tense to describe characters ...

T09: I see she's really good [of course] some minor grammatical issues, but I couldn't reject it.

T11: She said, 'His dad died when he was young' and that was the past tense. So, I didn't hear a lot of present tense and sort of language. She led the way a lot, I think.

This disagreement was considered unproductive because T11 just did not want to continue to discuss the issue any longer. What she did was indicate her disagreement to T12 and then switched to talk about this student's active engagement in the conversation with the other two male students.

Finally, in another example, T09 in most cases agreed with T11 about S2's talk. However, she disagreed with T11 regarding this student's use of tense. T09 thought that S2's limited talk time prevented her from being able to assess his actual ability in using the language. She did expect to see more from this student, and said, 'I didn't hear the

tense as much as I wanted to'. However, T11 observed the opposite. She commented that was satisfied with S2's use of tense and claimed, 'I think I could hear the tense, he was fairly consistent. However, neither of the two teachers attempted to convince each other, but simply expressed contradictory views, and made their assessment decisions based on these. Therefore, they did not receive any benefits from the discussion.

In identifying teachers with which decision-making style exhibited more disagreement than those with other styles, most of the disagreement was identified as belonging to those teachers with a self-regulated assessment decision-making style. All teachers have been so far identified to be comparable regarding background, biases and perception disagreement. Therefore, it is important to know how teachers with one decision-making style performed in comparison with the other two in terms of variability and consistency.

6.5.4. Variability and consistency differences among groups

6.5.4.1. Variability

Further analysis was conducted on actual assessments so that different decision-making processes somewhat resulted in different assessment decisions. Details on how individual teacher assessments were compared to one another have been previously presented in Chapter 4. Thus, this section only focuses on pointing out discrepancies between groups of teacher assessments. As can be seen in Table 6.2, comparisons of the means of actual scores assigned by teachers with each decision-making style for different student performance show discrepancies. For example, as for S1's performance, teachers with automated decision-making style were found to give the lowest scores to this student. The mean of actual scores by this group was 3.3, compared to the overall mean, in terms of variability, of 2.8. Meanwhile, teachers with self-

regulated decision-making style had the mean score of 2.5, meaning that these teachers assigned the highest scores to this student. The teacher with conflicted decision-making style tended to demonstrate the most variation in her score for S1, her assessment was significantly smaller the overall mean score at 2.0, indicating she gave the lowest score to this student.

Additionally, analyses also show that, as for variability for S2 assessments, overall the teachers with self-regulated decision-making style indicated the least variation and gave the better scores than those with conflicted and automated styles, with the mean score at 2.6 compared to the overall mean score of 2.71. The conflicted decision-making style teacher was identified to give the lowest score at 3.5, meaning that her assessment for this student accommodated the most variations. Assessments by teachers with automated decision-making style were a fraction higher than the overall mean score, 2.9 compared to 2.71, indicating that they were slightly stringent with this student. Regarding the last student, the overall mean score was 3.42 and teachers with automated decision-making style were found to give the best score to S3. Their mean score at 3.4 also means that their assessment accommodated the least variation. Giving a slightly higher score than the overall mean score, 3.5 compared to 3.42, the conflicted decision-making style teacher was more generous than those with the other styles. Finally, the mean score of teachers with self-regulated decision-making style was the lowest at 3.25, meaning that they were strict in scoring this student.

Table 6.2

Differences among Decision-making Styles: Variability and Consistency

Styles	Teachers	Variability			Consistency		
		S1	S2	S3	S1	S2	S3
Self-regulated	T03	4.0	3.0	3.0	0.55	0.68	0.58
	T05	2.0	3.0	3.5	0.76	0.87	0.39
	T06	2.5	2.5	3.5	0.43	0.27	0.42
	T07	2.5	3.0	3.5	0.57	0.49	0.51
	T11	2.0	2.0	3.0	0.67	0.70	0.51
	T12	3.0	2.0	3.0	0.74	0.75	0.56
	Mean	2.5	2.6	3.25	0.62	0.63	0.50
Conflicted	T04	2.0	3.0	3.5	0.64	0.42	0.54
Automated	T01	4.0	3.0	3.5	1.10	0.63	0.32
	T02	4.0	3.0	3.0	1.12	0.70	0.46
	T08	2.5	3.0	3.5	0.43	0.27	0.42
	T09	4.0	2.5	3.5	0.69	0.51	0.39
	T10	2.0	3.0	3.5	0.64	0.47	0.32
	Mean	3.3	2.9	3.4	0.80	0.63	0.38
	Overall mean	<u>2.80</u>	<u>2.71</u>	<u>3.42</u>	<u>0.69</u>	<u>0.61</u>	<u>0.45</u>

6.5.4.2. Consistency

Similar analyses were synthesised to compare differences about the ways in which teachers with three decision-making styles differed from each other in terms of consistency. To recap, consistency is the degree of difference or distance between the mean score and the scores assigned by teachers. It means that the smaller the difference is, the better the assessment was made. The result in Table 6.2 shows that teachers with automated decision-making style tended to produce the least consistent assessment for S1's performance, followed by the teacher with a conflicted assessment decision-

making style and teachers with a self-regulated style. For example, the difference between the average score of the automated style teachers assigned for S1's performance and the overall mean score by all 12 teachers was 0.80, followed by 0.64 and 0.62 for the conflicted and self-regulated style teachers, respectively. Clearly, teachers with self-regulated decision-making style assigned the most consistent scores when they assessed S1's oral output. As for S2's performance, the conflicted style teacher was identified as producing the most consistent assessment with the difference of 0.42 between her score and the mean score. Teachers with the conflicted and automated styles had the same degree of consistency in assessments for S2's performance, namely at 0.63. Finally, a different situation was observed among the three groups regarding consistency in assessing S3's output. In this situation, the automated style teachers were found to make the most consistent assessment at 0.38, while those from self-regulated and conflicted style groups were at the second and third position, namely at 0.50 and 0.54, respectively. Overall, the automated style teachers and the self-regulated teachers were the most consistent in their assessment across student performances, leaving the second teacher with the least consistent assessment decisions.

It is also worth examining the internal consistency within groups for consistency patterns. It can be seen from Table 6.2, that as one of the two most consistent decision-makers, the automated style teachers' consistency degree tended to improve after each time they assessed a student output. For example, their consistency for S1 was 0.80, this then reduced to .063 and 0.38 for S2 and S3 respectively. The self-regulated style teachers, despite having the same degree of overall consistency across students, demonstrated slight variations in their consistency. Their consistency degree was first 0.63 for S1, then rose to 0.63 for S2 before dropping to 0.50 for S3. The consistency

pattern of the conflicted style teacher was the most unstable and unpredictable when her consistency degree fluctuated at 0.64, 0.42 and 0.54 for S1, S2 and S3, respectively.

To summarise, teachers with different assessment decision-making styles demonstrated distinctive styles of assessment decision-making. There is some evidence that these decision-making styles may have interacted with differences in their backgrounds, assessment-related factors and their disagreements in evaluating student talk through moderation, resulting in differences in their assessments. Teachers with automated decision-making style had greater than those from the first group in terms of age and experience. However, teachers with automated style worked with students at higher levels than those with self-regulated decision-making style. In association with interactions with background and foreground factors, the automated decision-making style teachers' assessment tended to interact more with background factors than with assessment-related factors. Meanwhile, assessment decisions made by those teachers with the other two assessment styles were equal in interaction with both groups of factors. In terms of disagreement in perceiving student oral language outputs, the self-regulated style teachers were found to most readily adjust to and accommodate disagreements with other teachers. In relation to assessment variability for individual student performances, teachers from the self-regulated style group better fulfilled their jobs as classroom assessors than those from the conflicted and automated style groups. In addition, certain patterns were observed from the assessments of teachers with conflicted and automated decision-making styles across students. While the first group tended to be more and more gentle in their assessment, the third group's assessments fluctuated across students but always remained above the overall mean score. In terms of assessment consistency, the automated decision-making style teachers were one of the two most consistent decision-makers and their consistency tended to improve across

students. The self-regulated style teachers stably assigned consistent scores, whereas the conflicted decision-making style teacher was found to produce the least consistent assessments that were unstable across students.

These observations suggest that decisions made by teachers with conflicted and automated decision-making styles may not be as trustworthy as those made by the teachers with self-regulated decision-making style. Decisions made by the automated decision-making style teachers may be problematic, because they solely relied on their first perceptions about student talk. Because ‘perceptions are not reality; perceptions are filtered through the lens that we use to see reality’ (Anderson, 2003, p. 145), an initial assessment Gestalt is only one lens to judge language development. To make sound assessment decisions, teachers need to compare the information they have derive from their initial Gestalt with other available information sources (e.g., information obtained from consulting the assessment tools). As for decisions made by the teacher with the conflicted decision-making style, these were not dependable due to the teachers’ indecisiveness and uncertainty. Her decisions were made in the absence of reliable information, because her perceptions were unstable and vague. Like others, she formed her initial Gestalt about student output, but those first perceptions gradually faded when she engaged with the assessment tools. However, the tools did not really help her develop further firm perceptions about the students’ output. This lack of consistency of information in decision-making is more likely to result in poor decisions being made (Anderson, 2003). As suggested by the literature in large-scale testing to resolve unreliable ratings (McNamara, 1996, 2000; Weigle, 1994, 1998), this automated assessment decision-making style should be amenable to retraining or retaining.

6.6. Conclusion

This chapter has provided a detailed description of how teacher assessment decisions were formed in this study. Three different decision-making styles were identified and described in detail: 1) an automated decision-making style, 2) a self-regulated decision-making style and 3) a conflicted decision-making style. As a first step in making assessment decisions, teachers, irrespective of decision-making style, formed an initial assessment Gestalt that gave them general ideas about where the students were in terms of proficiency. However, the path the teachers then took in terms of their Gestalt decided their assessment decision-making styles. This study has found a tendency that assessment by teachers with an automated style and a conflicted style may not be as accurate and; therefore, less trustworthy because their decisions were formed with less consistent and reliable information. Conversely, assessments by those teachers with a more self-regulated style were more accurate and dependable. Comparisons among the three decision-making styles were conducted correlating demographic information, distribution of interactions, disagreement in perception and variability and consistency. Demographically, the results showed that teachers with one decision-making style were different from others in terms of age, experience and current teaching position. Teachers in self-regulated style group were older and more experienced than those in conflicted and automatic style groups. As for distribution of interactions, the automated decision-making style teachers were affected equally by background factors and assessment-related factors, whereas those teachers with the other two decision-making styles tended to interact more with assessment-related factors than with background-related factors. In terms of disagreement in perceptions of student outputs, self-regulated decision-making style teachers were the most accommodating of the disagreements among teachers. Finally, teachers with the three decision-making styles

were different from each other in that there was a tendency that self-regulated teachers were more consistent and less variable than those with these automated and conflicted styles in terms of variability and consistency.

The model of teacher assessment decision-making styles presented in this chapter revealed several issues regarding the process of teacher language assessment decision-making. These are: teachers' assessment practice, teacher assessment interactions and characteristics of different assessment decision-making styles. These issues are discussed in relation to the literature in the following chapter.

Chapter 7. Discussion

7.1. Introduction

This research study was conducted to investigate how experienced EAL/D teachers assessed the oral language development of students they did not know, using assessment tools that they were not familiar with. The aim of the study was operationalised into three research questions. The first examined teachers' assessment in terms of variability and consistency. The second identified interactions between teacher assessments and factors related to teacher background and the assessment process. Findings from the first two research questions shed light on identifying characteristics of teachers' assessment decision-making processes, helping to explain how teachers' assessment decisions were made.

The findings from these research questions have been presented in Chapters 4 to 6. In Chapter 4, teacher assessments were examined, and a wide range of differences were identified among teachers in terms of variability and consistency. It was also reported that there was a tendency for teachers to display specific assessment patterns across student performances and assessment categories. Findings on the factors which influenced teacher assessments were then presented in Chapter 5. It was reported that statistically, teacher assessment decisions were affected by some background factors such as age, qualification and the main language group taught. Teachers were also biased by several assessment categories in the criteria. In Chapter 6, a model was presented of the teachers' decision-making processes derived from the data. It was found that to make a sound and reliable assessment decision, teachers would normally go through three distinct stages of the decision-making process. Teachers in this study

were classified into three groups with three different decision-making styles. These were the self-regulated style, conflicted style and automated style. Only teachers with the self-regulated style participated in all the required stages of assessment decision-making.

The discussion in this chapter will be framed around three main themes arising from the findings of this study. The three major themes are: 1) a framework for better understanding teacher assessment decision-making processes, 2) the role of moderation in enhancing the trustworthiness of teacher assessment and 3) the sociocultural aspects of teacher assessment. The discussion of the framework for teacher assessment decision-making processes will contribute to better understanding of how teachers make judgement decisions and provide a model of teacher decision-making. The section on the role of moderation in enhancing trustworthiness of teacher assessment will highlight the importance of dialogic interactions, conflict and moderation in achieving assessment trustworthiness. Following this, the significance of understanding different sociocultural aspects of teacher assessment and engagement will be discussed to limit variability in assessing student development and to ensure consistency of teacher assessment decisions.

7.2. A New Approach to Understanding Teacher Assessment Decision-Making

It can be claimed from findings in this study that teacher assessment decision-making is about making judgements about the quality of specific performance samples and this process involves the application of a wide range of resources. This claim is well supported by Klenowski and Adie (2009). This study has found that the process by which a teacher assesses a student's output consists of three stages that differ from the

three sequential components proposed by Sadler (1998) including teacher attention drawn, teacher assessment using scoring rubric and teacher judgement decision. The teacher first pays attention to student performance, then assesses this performance against standards of some kind and, finally, makes a final judgment of student ability. The most significant finding of this study is that different teachers use different styles in their assessment decision-making processes. The styles described in Chapter 6 mark a new step in understanding in teacher-based language assessment, as, although several frameworks have been developed to highlight teachers decision-making processes to light, most of these are applied in the general area of decision-making (Klein, 1997), or general classroom assessment decision-making (McMillan, 2003), or in the assessment of writing (Davison, 2004). The decision-making model developed in this study is somewhat aligned with those previous ones, but more specifically describes a decision-making process for teacher assessment of spoken language. In this section of discussion, the framework by McMillan (2003) is revisited, followed by a justification of the way in which my proposed framework is in line with the framework, but is also significant in the discipline of language assessment.

Supported by Black and Wiliam (1998a), McMillan (2003) developed a framework to characterise teachers' classroom assessment decisions. Although this model is more likely to be used in general classroom assessment and grading contexts, its underlying principles can still be applicable in more specific teacher-based assessment settings. Accordingly, understanding teacher decision-making is to understand the relationships among five primary elements. These are 1) teacher knowledge, beliefs, expectations and values, 2) decision-making rationale, 3) external factors, 4) classroom realities and 5) assessment practices. These elements are interrelated, but 'the main tenet of these relationships was that assessment decision-

making was characterised by tension between the internal beliefs and values and external influences that are imposed on them' (McMillan, 2003, p. 35).

Accordingly, in McMillan's (2003) teacher classroom assessment decision-making process, teachers have their own knowledge, beliefs, expectations and values about assessing their students. These are also categorised into five influential themes including: 1) 'pulling for' students, 2) philosophy of education, 3) promoting students' understanding, 4) the need to vary assessment to accommodate diversity among students and 5) teacher motivation to enhance student active engagement. However, the purpose of classroom assessment for teachers is not only learning and teaching support—they experience tension created by external factors as assessment is also for accountability, policies and reporting to parents. Further, there are other elements that are beyond the control of teachers that they must confront when making their decisions.

From what was found in this study, I would suggest that to better understand teacher decision-making process when assessing spoken language, importance should be attached to understanding sources of variability in teachers' assessment decisions, the role of teacher flexibility and metacognition in assessing student oral work, and the role of scoring rubrics in improving variability and consistency in teachers' assessment decision-making.

7.2.1. Using variability as a resource in teachers' assessment decisions

As noted by experts in the field of language assessment (Davison, 2004; Davison & Leung, 2009; McNamara, 1996), variability is an inherent characteristic of assessors as humans. Therefore, it is important to understand that it is impossible to fully eliminate variability among teachers; in fact, in an assessment for learning paradigm variability is the foundation for developing trustworthiness, as it provides the

basis for robust conversations and interactions about students development which can improve teacher understanding of their own decision-making processes and biases, leading to greater internal consistence, and paradoxically, less variability.

In relation to variability in teachers' assessment decisions, this study found that variability is caused by a range of factors, including any variables directly or indirectly related to assessment tasks, assessors or the entire assessment process. This is in line with other previous studies (see Barkaoui, 2010c; Eckes, 2005, 2008; Leckie & Baird, 2011). At the same time, this study has provided new insights into ways to understand and even exploit the sources of teacher assessment variability. As the framework of teacher assessment decision process suggests, teachers may consciously consider several different elements when they made their judgements. These elements are classified into two categories: assessment-related factors and background-related factors. Assessment-related factors include knowledge about students such as gender (Carroll, 1991; Eckes, 2005; Lumley & O'Sullivan, 2005; Porter & Hang, 1991), accent (Carey et al., 2011; Cargile & Giles, 1998; Edwards, 1982; Gass & Varonis, 1984; Gill, 1994; Major et al., 2002), assessment task (Fayer & Krasinski, 1987; Kim, 2009; Lumley & McNamara, 1995; Luoma, 2004; McNamara, 1996; Weigle, 1998, 2002) and assessment criteria (Lumley, 2002a; Rezaei & Lovorn, 2010). In addition, this study also found that some teachers were aware, while some other were not, that their assessment decisions may have been affected by their current teaching position, English teaching qualification, their main exposure to students from a language background and students' personalities.

Knowledge of factors that influence teacher decision-making helps build a better understanding of what teachers consciously and subconsciously may consider when assessing their own students' oral work and helps support teachers to reflect on their

own assessment processes and how to improve their trustworthiness, and through that, their classroom assessment practice. When in their roles as classroom assessors, teachers tend to be influenced by factors that they do not think they are affected by. For example, in this study although quantitative analysis indicated that teachers' age somewhat influences the way in which teachers make their final decisions, the teachers themselves were not aware of the existence of this influence. At other times, teachers' assessments were influenced by factors that they had been aware. However, they did not or could not prevent this from happening. For example, teachers were reportedly affected by several variables that they knew would have certain effects on their assessment decisions. Those factors include teacher experience, qualification, main exposure to a language group, assessment criteria, assessment task, assessment procedures and student personality.

Therefore, I would suggest that identifying the variables of teacher assessment decision-making will lay a strong foundation for improving teacher assessment practice and ensuring the trustworthiness of classroom assessment. To use variability in teacher assessment productively, it is important that all factors causing such variability should be understood.

Given that variability is an inherent characteristic of human assessors and that removing it is not possible (Davison & Leung, 2009; McNamara, 1996), assessment trustworthiness can be better ensured by making variability in teacher assessment more transparent and explicit. All information on influential factors of teacher decision-making can be used as a source of reference for teachers and schools to help better understand the multiple influences on teacher assessment decisions. If communicated to teachers, this kind of information, instead of being seen as implicitly judging and criticising their assessments, will help them enhance their assessment literacy and

practice. Such communication enables teachers to be more aware of and avoid or at least minimise the effects of such factors on their assessment practice. For example, in designing assessment tasks, developing scoring rubrics, administering assessment tasks and assessing students' responses, teachers will be more cautious about what may affect their decisions and try to be compensate for their biases to ensure their decisions are dependable. In addition, information on influences in teacher assessment decision-making can also be useful for teacher education. Teacher educators can make this information known to pre-service teachers by embedding it into training material. Overall, identifying and understanding what shapes teachers' assessment decisions can improve the quality of their assessment practice.

7.2.2. The role of flexibility and scoring rubrics in improving consistency in teacher assessment decisions

As can be seen from the framework of teacher assessment decision-making, teachers from three groups use three different decision-making styles, although there may be more. Quantitative results indicate that teachers with a self-regulated assessment decision-making style are more consistent and less variable than those with a conflicted assessment decision-making style and those with an automated assessment decision-making style. As described in Figure 6.1, only teachers with a self-regulated assessment decision-making style experienced the three stages in the assessment decision-making process. Therefore, during discussions about teacher assessment decision-making, the process of decision-making of this group of teachers will be used to exemplify arguments about the matter.

7.2.2.1. Role of flexibility

From the findings of this study, I suggest that flexibility is one of the integral components of consistency in teacher decision-making. Flexibility, which involves self-reflection and willingness to change, can partly explain why teachers with self-regulated assessment decision-making styles outperformed teachers from the other two groups. This study showed that at the beginning of the decision-making process, teachers develop their initial perceptions about student performances when they listen to them. More consistent and reliable decisions can be made with more reliable information (Anderson, 2003). Initial perceptions are a single lens to view student ability (McMillan, 2003) and additional lenses are needed, with teachers able to identify the contradictions between their initial perceptions and subsequent evidence and be willing to change their judgments. As presented in the framework of teacher assessment decision-making in Chapter 6, flexibility is one of the contributors to the sound decision-making of self-regulated teachers, whose assessments were more consistent and trustworthy than those of the other two groups. Flexibility comes from actual experience through assessment and teaching practice in classroom contexts and from active engagement with assessment tools, assessment criteria or rubrics. It is worth noting that flexibility in teacher decision-making is a socio-cognitive process in which professional knowledge plays an important part. This finding is supported by Colton and Sparks-Langer (1993). Notably, their professional knowledge base influences teachers' interpretation of what is to be assessed. Demonstrating flexibility, teachers consider the content of student output, what they may have been taught and what they were asked to do in a speaking task. Teachers then consider their knowledge of the students to be assessed. This kind of understanding about students' demographic and cultural background helps teachers to better decide if a pattern of student language

performance is their weak point or a common phenomenon in language use among speakers of their cultural background. Furthermore, the role of the teachers' prior experience in building flexibility is important (also see Kennedy, 1989). Due to individual differences among teachers, prior experiences also vary. Flexibility helps teachers to link what they are seeing or hearing to what they have experienced. Challenging and examining their initial perceptions through flexibility leads to a better course of action being adopted (Colton & Sparks-Langer, 1993).

Flexibility is also built through teachers' consideration of context of the performance task. When making an initial impression of a student's ability to communicate, particularly when it is a negative impression, a good assessor will seek reasons to explain why the student has communicated in such a way. Teachers may reflect about many contextual factors including how the task is carried out, what time of the day it is conducted, whether there are any sources of distraction such as presence of the camera filming them, conversing with others, or whether their peers affect the way in which they use the language. As an example, some teachers realised S3 did not seem confident because he often failed to maintain eye contact with the teacher interviewer. The teacher belief of wanting to 'pull for students' proposed by McMillan (2003) fits this situation. That is, the teachers sought for and were satisfied with an explanation that a lack of eye contact could be because that student came from a cultural background in which a young person is not allowed to look straight into a senior's eyes. Flexibility around contextual factors does not mean that students will be treated more favourably, instead, those factors help the teacher to consolidate and validate their initial Gestalt to form a strong basis for more consistent decision-making.

Finally, this study found that teachers tend to include their personal and social values in the assessment process. These values are developed by 'one's family, personal

encounters, reading and life experiences' (Colton & Sparks-Langer, 1993, p. 47). The more flexible teachers demonstrated a belief and willingness to do whatever they can to help students be successful (McMillan, 2003)... Through experiencing difficulties, even failure, in life, such teachers tended to view student weaknesses as a temporary phenomenon that occurs due to the effect of different contextual factors. Hence, these social values play a significant part in teachers' flexibility and influence their daily teaching and assessment decision-making practices.

Overall, flexibility provides teachers with more opportunities to examine and re-examine from different perspectives what they believe to be the ability of students. These opportunities enable teachers to revisit their judgements and provide them with a firsthand understanding of what student can do in terms of using the language orally. Flexibility is necessary and important, irrespective of whether these first judgements are right or wrong. If the judgements are correct in the first place, flexibility will function as a dual facilitator. It validates teachers' first perceptions and, at the same time, raises teachers' confidence levels. Teachers become more confident with their assessment competence to make better assessment decisions. In cases in which the judgements are partly not correct or are wrong due to a deficit of information, flexibility gives teachers the chance to look for more reliable and consistent information, as well as to review the evidence and evaluate what is to be assessed (i.e., student talk) from another perspective before arriving at a final decision.

7.2.2.2. Role of scoring rubrics

In addition to flexibility, this study also found that teachers' engagement with assessment criteria or rubrics significantly impacts consistent and trustworthy assessment decisions. Scoring rubrics offer many benefits to assessment and instruction

(Brookhart, 2013; Brookhart & Chen, 2015) in assessments practice in general and in classroom assessment in particular. One of the benefits is enhancing consistent scoring in teacher assessment, especially if they are analytic and topic-oriented rubrics illustrated with exemplars (Barkaoui, 2010c). In this study, one of the differences between the self-regulated decision-making style teachers and the conflicted and automated decision-making style teachers is their engagement with the scoring rubrics. Engaging with assessment rubrics to evaluate student learning enables teachers to reflect on how the instruction has been designed and implemented and how learning has been organised. Such reflection helps teachers decide on what should be done to improve teaching and learning.

This study found that there are two main categories of rubrics, one mental and highly individual, constructed by teachers in their heads, and the other common and concrete, given to them as an assessment tool to score the assessment task. When assessing their students' talk, teachers seem to employ both kinds of rubrics, even though they are supposed to follow the common one. That teachers develop their initial Gestalt—an overall perception about the ability to use the language of students when they first listen to student performances—indicates that subconsciously they use a holistic scoring rubric. Such a holistic rubric is probably developed immediately prior to the commencement of the assessment process. This instant holistic scoring rubric is generated based on teachers' prior professional knowledge, experience and expectations. In association with the former, when teachers listen to students responding to the speaking tasks, what they see for the first time provides them with a general understanding about the students' ability in using the language. Thus, teachers' mental schemas or rubrics help them initially place students on the proficiency continuum. Through this, the inference can be drawn that those teachers with more experience may

outperform novices in developing an instant holistic rubric (Barkaoui, 2010c). It may be also inferred that it is less time-consuming and more manageable for experienced teachers to generate and use these rubrics than it is for inexperienced ones who tend to benefit more from analytical scoring rubrics.

The findings of this study suggest analytical scoring rubrics help teachers to focus their attention more closely on the various criteria for scoring and, therefore, improve internal consistency in teacher assessment. These findings are consistent with Barkaoui (2010c) and Barkaoui (2007). The main difference between the self-regulated decision-making style teachers who were reported to produce consistent and trustworthy assessment decisions and the conflicted and automatic decision-making style teachers is the level of engagement with the assessment tools (e.g., the analytical scoring rubric). All teachers start their decision-making process by developing their initial Gestalt about students' overall proficiency based on what their first impressions. Being aware that their initial Gestalt may not be sufficient to rely on, self-regulated decision-making style teachers pay significant attention to the scoring rubrics to compare with their firsthand observations about student performance, before final decisions are made about the student. Apart from flexibility, engagement with the assessment tools seem to be an appropriate and necessary approach to revisiting initial observations, even though it may be time-consuming (Mertler, 2001; Nitko, 2001) or prevent teachers from seeing students' creativity in their responses (Wolf & Stevens, 2007).

In fact, this study showed that engagement with the scoring rubrics helps teachers in several ways. First, scoring rubric engagement gives teachers an opportunity to test their initial Gestalt about student overall proficiency. An analytical rubric like the one in this study has been proven to focus most teachers' attention on its scoring categories instead of student performance as a holistic rubric does (Barkaoui, 2010c). Its

specific scoring categories provides teachers with a means to perceive student performances in different language areas and, simultaneously, to triangulate their first perceptions. While their first perceptions give them just a general sketch of what students can do, the specific scoring categories depict a full and detailed picture about students' performances in terms of linguistic and communication competences. The role of scoring rubrics in this first instance; therefore, is important no matter whether the initial Gestalt is right or wrong. In the first case, if the initial Gestalt and the result generated by engagement with the rubrics are similar, this strengthens teachers' confidence in assessing students' work. This is important for teachers in making assessment for learning purposes (Gears, 2005).

This study shows another benefit of engagement with scoring rubrics comprising specific categories, that is, it helps to anchor teachers' assessments (Wolf & Stevens, 2007). When engaging with scoring rubrics, the teachers' attention was constantly drawn to the specific scoring categories. What they do is try to interpret different areas of student performances and match these with proficiency indicators in equivalent scoring categories. In doing this, teachers tend to stabilise their application of the rubrics across students. Thus, by applying scoring rubrics that are response-focused or task-specific, teachers are more likely to improve their own consistency when assessing different students performing on the same task. In classroom assessment contexts, a higher level of internal consistency results in fewer concerns about student bias (Wolf & Stevens, 2007, p. 12).

Finally, this study shows that fairness in teacher assessment is more likely to be achieved with teacher engagement with scoring rubrics. Fairness in classroom- or - teacher-based assessment means a student responding to a speaking task should receive a similar score no matter how many times they are assessed by a teacher or how many

teachers assess them. Hence, fairness is referred to as self-consistency in the former situation and as cross-consistency in the latter situation. Since teachers are human, they exhibit a range of individual differences in their decision-making (Baker, 2012); therefore, variations in their assessment decisions are unavoidable (see Davison & Leung, 2009). This study shows that rubrics assist by providing teachers with all relevant information, to enable teachers with an insufficient amount of information retrieved from the stimulus input (i.e., student responses) to compare their findings. Such rubrics also help teachers to mitigate the effects of their professional experiences and personal backgrounds when processing and evaluating information to make final judgements.

Overall, the findings in this study show that scoring rubrics make a significant contribution in many ways to the assurance of consistency in the process of teacher assessment decision-making.

7.3. Moderation to Improve Consistency

The findings of this study reinforces the role of moderation in building consistency in teacher assessments and in enhancing the trustworthiness of the overall assessment system (Maxwell, 2010; Hipkins, 2010a). The findings of this study provide more insights into how moderation contributes to consistency in teacher assessment. In this study, most teachers found the moderation following each assessment section very important and helpful in enabling them to share their perceptions and ideas. That teachers support moderation confirms the importance of moderation in improving the quality of their assessment. This echoes the positive attitudes towards moderation in the literature (see Connolly et al., 2012). Although one teacher in the study found moderation frustrating, because teachers in her group did not agree with her and she felt this

reduced her confidence in her decision-making, this teacher subconsciously gained a great deal from this discussion when she had a chance to reflect later. This reinforces the argument that trustworthiness in teacher assessment decision-making is more likely to be reached by teachers expressing their disagreement followed by justifying their opinion, than from immediate agreement (Davison, 2004).

In conclusion, this study shows that moderation, as a process of sharing, reviewing and reconceptualising teachers' understandings, is necessary and important in ensuring consistency in the teacher assessment decision-making process, and can provide even more benefits for teachers who hold different perspectives.

7.4. Revisiting Trustworthiness

Since the general purpose of this study is to improve the trustworthiness of teacher assessment systems in classroom contexts, it is necessary to revisit this concept taking into account the findings of this study. As referred earlier, trustworthiness is usually equated with reliability, a concept popularised in psychometric assessment perspectives. It refers to the degree of agreement between different raters (i.e., inter-rater reliability) and within the same rater (i.e., intra-rater reliability) (see Gamaroff, 2000; McNamara, 1996, 2000; Weigle, 2002). Due to the nature of psychometric assessment and the importance of high-stakes tests, it is imperative to achieve reliability in assessment decisions. Failure to do so would result in raters being trained or retrained or even removed from the assessment process (McNamara, 1996).

However, trustworthiness, from the standpoint of this study, is more about the process through which teachers are able to share their disagreements, justify their opinions and arrive at a mutual understanding of student performance standards (Davison, 2004; . Davison & Leung, 2009). From this classroometric perspective

(Brookhart, 2003), differences among teachers in terms of perceptions and interpretations of student output and assessment standards provide the starting point and stimulus for teachers to engage in professional conversations through a process of moderation. The findings of this study show that absolute agreements do not always mean that teachers have had the chance to reflect on the process of decision-making and their decisions.

An important ingredient for trustworthiness is the opportunity for teachers to make explicit and justify opinions (Klenowski and Adie, 2009). This should be considered one of the crucial components of teacher professional learning as it allows teachers to view and understand student performance from another person's perspective, and then to accommodate to that perspective or not, as part of the moderation process.

Conclusion

This chapter has presented a detailed discussion of the contribution of the findings of this study to the literature on variability in oral language assessment decision-making process in classroom contexts. The process of teacher assessment decision-making is a complex process accounted for by the interplay of many influential factors that are related but not limited to assessment contexts and backgrounds and involve different assessment pathways. The quality of teachers' assessment decisions is dependent on three significant contributors (e.g., teacher flexibility, rubrics and moderation). Taking all these elements into account can build a more trustworthy assessment system – the implications of this for policy, professional learning and practice will be discussed in the next chapter.

Chapter 8. Conclusions and Implications

8.1. Introduction

This study has explored variability in the decision-making of experienced EAL/D teachers in NSW when using unfamiliar assessment tools and materials to assess performances by unfamiliar students. The aims of this study were threefold: (1) to explore how consistent teacher assessments were, (2) to identify the factors that shaped teacher assessment decisions and (3) to provide more insight into the process of teacher decision-making. The study was conducted in two stages. The first stage involved a questionnaire to collect teachers' demographic data, assessment tasks to collect information on teachers' assessments (scores) and moderation. The second stage involved a retrospective think-aloud that enabled teachers to justify their assessment decisions, followed by interviews with individual teachers to clarify their justifications and to collect more information on their decisions.

In Chapter 4, findings on teacher assessments in terms of variability and consistency were presented and discussed along different dimensions. Chapter 5 provided a detailed description of how teacher assessments were affected by factors that related to teachers' background and assessment contexts. In Chapter 6, the different decision-making styles used by teachers to support their assessment decision-making were described and a framework for teacher assessment decision-making presented. Significant findings from the study and their contribution to the literature were discussed in Chapter 7. In this chapter, a summary of the key findings of this study will be first presented, followed by their implications. The limitations of the study will be outlined and suggestions for further research presented.

8.2. Summary of Key Findings

Several main findings about EAL/D assessment practices in NSW were revealed through this study. The key findings from the assessment tasks indicate that there were significant differences in teachers' scorings in terms of both variability and consistency. The differences were also reported to occur not only in overall assessments, but also for individual students and assessment categories. Common tendencies were observed in that, as a group, teachers tended to be more tolerant and consistent in assessing one student's talk but more stringent and inconsistent in assessing other students' talk. A similar tendency was also identified for assessment categories. These key findings indicate that teachers may have observed and considered beyond-performance characteristics of students in making their judgements. The findings also suggest that teachers may have constructed and applied their own individual standards in decision-making as a substitution for, or as a complement to, the set of common standards that were given.

Further findings indicate that teacher assessments were both consciously and subconsciously influenced by several factors in association with teachers' backgrounds and the assessment contexts. Thus, teachers may or may not have been aware of the actual interaction between their assessments and influential factors. This suggests that in making assessment decisions, teachers not only focused on what they were supposed to assess, but they also placed importance on elements other than what students could do. These findings can be used to partly explain the differences in teacher assessment decisions.

The final, but most significant, key findings reveal insights into the teacher assessment decision-making processes and assessment styles, and a framework to understand how teachers make assessment decisions. Quality assessment decision-

making is seen as a self-regulated socio-cognitive process involving three stages. In the first stage, teachers developed their initial assessment gestalt about student performance, based on their impressions of student strengths or weaknesses. This Gestalt gave them general ideas of where they would place students on the performance continuum. In the second stage, teachers revisit their Gestalt through a process of flexibility or reflection and engagement with the scoring rubrics that either confirmed the Gestalt or indicated that it would need to be adjusted. In most cases, modifications were required. Following the final stage, teachers made their assessment decisions. These assessment styles, may be dynamic and context-specific. One can even question if these assessment style will change according to the context. This issue deserves further investigation.

As suggested in the framework, more weight needs to be given to the importance attached to the roles of flexibility, scoring rubrics and moderation in ensuring consistency in the teacher assessment decision-making process. Flexibility is gives teachers the chance to reflect on their initially developed Gestalt about student performances. Furthermore, engagement with scoring rubrics allows teachers to gather sufficient relevant information and evidence to arrive at sound decisions. The use of moderation is also highlighted. A shared understanding and opportunity for productive disagreements among teachers is a key component in developing greater consistency in teacher assessment decisions.

From these findings, various implications can be drawn regarding enhancing teacher assessment literacy, practice and learning support.

8.3. Implications

The implications drawn from this study include implications for theory and practice, implications for teacher training and implications for educational policy.

8.3.1. Implications for theory and practice

Findings from this study on teacher assessment decision-making provide a framework to revisit several key concepts, principles and practices in the field.

From a theoretical standpoint, there is a need to conceptualise teacher assessment and its implementation as a continuous iterative staged process which involves teachers in collecting, interpreting and evaluating information, at the same time building professional knowledge, flexibility and self-regulation. Information evaluation needs to be done alongside engagement with assessment standards. To ensure the consistency and trustworthiness of their decisions teachers also need to participate in ongoing assessment conversations through moderation in which productive disagreements are scaffolded and supported.

From a practical standpoint, teacher awareness needs to be involved in effective teacher assessment decision-making. Research shows that teachers' assessment decisions are affected by both internal and external elements. As a central agent of assessment, teachers need to understand what these are and why and how these have an effect. Resources should be developed by gathering and synthesising findings from a rich body of research to provide teachers with relevant knowledge. This would be helpful, as teachers would benefit from this resource instead of feeling criticised. Teachers' keen awareness of such concerns would help them avoid potential biases in making assessment decisions in their classrooms. At the same time, teachers would implement their assessments through careful reflection and more engagement with scoring rubrics and other available assessment tools, leading to greater flexibility and self-regulation. In practice, teachers' awareness of the various influences on their assessment should be operationalised into designing tasks and constructing assessment criteria to assess their students and supported through various forms of moderation.

As noted by Maxwell (2006), this study has revealed that moderation is not a passive process in which teachers simply talk about how much their judgements are in agreement. Rather, this moderation is an active and contested conversation involving teachers in personal comparison and alignment of their judgements. Thus, rather than acknowledging and accepting other viewpoints, or simply aiming for consensus, it is important to resolve differences in opinion.

One of the implications of this study is teachers need guidance and training in moderation. They need engage in a process of reconceptualisation including either deconstructing, reconstructing or co-constructing assessment conceptions (Stoll & Bolam, 2005). This means that teachers decide what to do with their own judgements to reach final agreement with one another.

In deconstructing, teachers bring to the moderation process their personal perceptions of student work standards, what Klenowski and Adie (2009) describe as the result of 'a complex interplay of many personal and professional referents besides the stated standards with some at times being more prominent in the decision-making process than others' (Klenowski & Adie, 2009, p. 20). A simplified version of this description, I would suggest, is that those perceptions are teachers' decisions about students' work prior to moderation. After the initial professional conversations with others in their moderation group identifies inconsistencies with others, some teachers decide to abort their own perceptions and knowledge of student outputs developed prior to meet with others. This only happens when they find that, after listening to others, their judgements are different and they are unable to justify their decisions. Rather they may find others' arguments significant and convincing. Therefore, they may find themselves in a situation in which it is necessary for their conceptions to be deconstructed to make way for a consensus.

The second possibility of reconceptualising is reconstructing conceptions. Contrasting their own conceptions about students' work results in the dissolution of their conceptions. Teachers cognitively feel a need to have new knowledge about students' works reconstructed. Reconstructing is understood in two possibilities. In the first, teachers may find that their perceptions are inconsistent with those of others, but they choose to preserve some of the thoughts they believe are true. Reconstructing in this regard involves making slight modifications to their previous conceptions. Modifications may be adapted from perspectives of other teachers. In the second possibility, this cognitive activity may involve teachers compromising their previous thoughts to admit and then accept the others' perspectives.

Co-constructing is the last step in the reconceptualising process. This step refers to the extent to which negotiation is the nature of teachers' professional conversations. For example, if three teachers in a moderation group each have a different perception about a student's output, the focus of social or calibration moderation should always be on and, for a student's performance (Klenowski & Adie, 2009; Timperley, 2008), and the teachers willing and open to negotiation. They negotiate by presenting their arguments in defence of their ideas. This may allow teachers to capture new elements of a student's work that they have not observed. A teacher's perception about a student's work is like viewing the world through a single lens; therefore, three views are better than one. What one teacher cannot see can be identified by another. If they can construct a new perception about a student's work together, the quality of the assessment can be improved considerably. Thus, differences in perceptions are seen as normal because there is always individual differences in teachers' assessment decision-making. Of course, moderation can only be considered successful when teachers discuss and

negotiate to complement each other rather than to compete for each other's attention (Hipkins & Robertson, 2011).

In addition to moderation's immediate focus on making consistent assessment decisions, it is also useful to promote professional development among teachers (Hipkin, 2010a). Moderation and improving student achievement are causally related. First, as mentioned earlier, the primary focus of moderation is to ensure more trustworthy teacher assessments that, when communicated to students, can be a source of informed feedback to improve student outcomes. Further, through moderation, this assessment information can also be used as feed-forward to help teachers self-adjust and improve their teaching practice to improve student success. So moderation between teachers within schools or across a cluster of schools is important as a professional learning activity for teachers and can contribute to improving student outcomes (Carless, 2015).

8.3.2. Implications for pre-service teacher training and professional learning

The findings of this study about teacher variability and consistency, interactions in teacher assessment and characteristics of teacher decision-making also suggest implications for teacher training and professional learning.

First, as reported in this study, the participating teachers performed differently when assessing students' oral language samples. It is noted that the participating teachers were experienced in teaching EAL/D but were unfamiliar with the assessment tools, materials and students. Therefore, it is implied that to enhance teacher assessment literacy and trustworthiness of the whole assessment system, teachers need to be formally trained to use new assessment tools regardless of how experienced the teachers might be. As in the literature, this study found that assessment tools such as scoring

rubrics have a strong effect on how teachers assess student work. Saying this does not necessarily mean that teacher differences can be avoided solely through training teachers how to use assessment tools. However, better use of assessment tools would help teachers reach a consensus in their perceptions of student work, resulting in the improved chance of consistency in their assessment decision-making.

Another implication is that teacher educators need to consider individual differences among teachers when they develop professional learning programs. These individual differences are not limited to teacher demographics or backgrounds, but they can involve professional differences such as working experiences or working styles, and cognitive discrepancies. Teachers, their needs and their goals are not standardised, so the same piece of advice may not work for all teachers (McMillan, 2003). However, the findings on the process of teacher assessment decision-making also showed that teachers appear to have different assessment pathways with different assessment decision-making styles. Teachers who were more self-regulated decision-making completed three stages of decision-making, and in the process revealed that their assessment decisions were more consistent and dependable. These teachers were also involved in reflection and assessment tool engagement before they decided on their judgements. In contrast, many teachers did not pay sufficient attention to the scoring rubrics or skipped some stage of the decision-making process. This suggests teacher educators should aim to develop greater self-regulation and flexibility among teachers.

Given the important roles of flexibility and scoring rubrics in improving consistency in teachers' assessment decision-making, it is important to strengthen teachers' understanding and interpretations of scoring rubrics. To achieve this, formal training in designing and using scoring rubrics must be provided to teachers no matter

how experienced they are in teaching or assessment. They should also be encouraged to apply rubrics in their classrooms on a regular basis.

Finally, both pre-service and in-service teachers, apart from pedagogical knowledge and skills, also need to reach a deep understanding of what assessment is and why, how and when it is implemented. In addition to being trained to interpret and accommodate aspects of student performance into the scoring rubrics, it is also important for teachers to know how to construct such rubrics. In doing so, they can use the scoring rubrics more consistently and effectively. To avoid differences in the interpretation and application of scoring rubrics, teachers need to be trained to construct the rubrics explicitly. Performance descriptors in the rubrics need to be understood in the same way by all teachers. Thus, an explicit scoring rubric is more likely to result in a greater uniformity in understanding and interpretation.

8.3.3. Implications for educational policy

The findings of this study also reveal some suggestions for educational authorities or organisations. As mentioned in Chapter 1, this study was part of TEAL, the state-wide assessment project in Victoria that aims to enhance assessment literacy for EAL/D teachers. Considering the dramatic increase in the number of immigrant students in Australia and the lack of resources available for EAL/D teachers to improve their assessment practice, the project's reach should be extended throughout the country. It is important that all teachers are also provided with the required information and knowledge they need to make the most use of such resources, so that they can consistently use the materials to gradually improve their assessments as well as teaching practices.

8.4. Limitations and Suggestions for Further Research

This study aimed to investigate variability in teacher assessments and analyse the process of teacher assessment decision-making. Given that assessment aims to improve learning, a better understanding of how assessment decisions are made will facilitate better student learning outcomes. This study took one approach to understanding teacher assessment decisions; however, this may not be sufficient to identify all aspects of teacher assessment. Therefore, further studies need to be conducted to provide more insight into those aspects. Suggestions for further research are proposed based on some of the necessary limitations of this study.

The first suggestion for further research is to increase the sample size to test the generalisability of findings, specifically those on statistical differences in teacher scorings, interactions and the decision-making process. The outcome of this study is limited by the small sizes of the three samples namely teacher-participants, assessment tasks and students' sample performances.

Second, further research attention should be drawn to investigating the roles of teacher gender as well as other professional background-related factors in ensuring consistency and trustworthiness in assessment decision-making. Participating in this study was completely voluntary and only female teachers agreed to take part in providing information. Although the gender effect among teachers on their assessment decisions has not been well documented, the effect of gender on decision-making has been confirmed in other fields rather than classroom or teacher assessment (Chung, 2002; Frederick, 2005; Han, Hsu, & Lee, 2009; Harris, Jenkins, & Glaser, 2006; Johnson & Powell, 1994; Mitchell & Walsh, 2004; Powell & Ansic, 1997; Roxas & Stoneback, 2004; Venkatesh, Morris, & Ackerman, 2000). These suggest that men and women might be cognitively different in perceiving, processing and making decisions.

Another suggestion for further research involves the development of moderation guidelines. This study did not provide teachers with a set of formally developed guidelines for moderation. This may have resulted in the unproductive disagreements of some teachers who were viewed as having the automated assessment decision-making style. However, as in the process of assessment, teachers may differ from each another in moderation styles as well as moderation decisions. Therefore, research studies focusing on moderation will help provide greater insight into the role of moderation in helping to improve consistency in teacher decisions. There may be a need to reconceptualise moderation in ways that teacher styles, needs and expectations can be addressed and accommodated.

Further research attention should also be paid to finding a better think-aloud or developing more effective tools to obtain more insights on socio-cognitive processes. In this study, retrospective think-aloud protocol was employed to give teachers opportunities to justify what they did. Retrospective verbal protocols have been reported to be less problematic than concurrent verbal protocols, allowing assessors to fully concentrate on their work (Bowers & Snyder, 1990; Van Den Haak et al., 2003). Yet it may hinder teachers from recalling what they did. Despite being a burden to teachers who must assess and verbalise at the same time, concurrent verbal protocol enables teachers to provide more articulation and comments about a student's work and what is happening in their head, compared to retrospective protocol. This is because verbalising while assessing means teachers can reflect from recent memory and; therefore, the information is more likely to be accurate and precise. Some of the teachers in this study indicated that they did not remember why or how they formed their assessment decisions because they had been involved a range of tasks during the interval. Although these concerns had already been foreseen and teachers were able to observe student

samples and their scorings again, information loss due to memory issues should still be considered. A research tool that avoids all the drawbacks of concurrent and retrospective think-aloud protocols needs to be developed.

Lastly, this study only dealt with one kind of teachers' assessment but did not deal with teachers' assessment contextualised in real classrooms. In reality, teachers have much knowledge about their students' performance and abilities, and such knowledge and information may potentially influence their decision-making processes. For this reason, further research in real classrooms should be done in order to see if teacher assessment decision-making styles change in different contexts.

8.5. Final Thoughts

This research has shed light on the process of teacher decision-making in an Australian context in which English is taught as an additional language or dialect. Variability in teacher assessment was explored and the process of how teachers made their judgements was investigated. The main contribution of this study is the suggestion of a framework for decision-making in teacher judgements of oral language outputs that places more weight on the importance of flexibility, scoring rubrics and moderation in ensuring consistency and trustworthiness in teacher assessment decision-making. This study also unearthed several factors that have an effect on the way teachers assessed students' oral communication. Overall, the findings of this study suggest that variability in teacher assessment decision-making should not be considered so problematic as it can be used to provide opportunities to improve the quality of teachers' judgements in language education.

Above all, this study provides a better understanding of teacher decision-making and alternative perspectives on variability in language assessment. Teachers are the lead

actors in all teaching and assessment practices that help students succeed in school.

Therefore, greater attention should be paid to teachers' current and future professional development. It is hoped that this study will help to raise awareness of current concerns and potential needs.

References

- The Australian Curriculum, Assessment and Reporting Authority. (2014). *English as an Additional Language or Dialect Teacher Resource*. Retrieved from http://docs.acara.edu.au/resources/EALD_Overview_and_Advice_revised_February_2014.pdf
- Abikoff, H., Courtney, M., Pelham, W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology*, 21(5), 519–533. <https://doi.org/10.1007/BF00916317>.
- Ahour, T., & Mukundan, J. (2009). Analytic assessment of writing: Diagnosing areas of strength and weakness in the writing of TESL undergraduate students. *Iranian Journal of Language Studies*, 3(2), 196–208.
- Akbari, R. (2012). Validity in language testing. In C Coombe, P Davidson, B O'Sullivan, & S Stoyhoff (Eds.), *The Cambridge guide to second language assessment* (pp. 30–36). Cambridge: Cambridge University Press.
- Aliaga, M., & Gunderson, B. (2000). *Interactive statistics*. Saddle River, NJ: Prentice Hall.
- Anderson, L. (2003). *Classroom assessment: Enhancing the quality of teacher decision making*. New York, NY: Routledge.
- Anderson, R. (1998). Why talk about different ways to grade? The shift from traditional assessment to alternative assessment. *New Directions for Teaching and Learning*, 74, 5–16. <https://doi.org/10.1002/tl.7401>.
- Angelo, D. (2013). NAPLAN implementation: Implications for classroom learning and teaching, with recommendations for improvement. *TESOL in Context*, 23(1/2), 53–73. Retrieved from <https://search.informit.com.au/fullText;dn=885422716767620;res=IELAPA>

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers* (2nd ed.), San Francisco, CA: Jossey-Bass.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999). *Standards for educational and psychological testing*, The University of Michigan: American Educational Research Association.

AusVELS. (2013). *English as an additional language (EAL) companion to AusVELS: For implementation in 2013*. Retrieved from http://www.vcaa.vic.edu.au/Documents/viccurric/eal/EAL_companion_to_AusVELS.pdf

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12(2), 238–257. <https://doi.org/10.1177/026553229501200206>

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly*, 9(3), 225–248. <https://doi.org/10.1080/15434303.2011.637262>

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12(2), 86–107. <https://doi.org/10.1016/j.asw.2007.07.001>

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, 44(1), 31–57. <https://doi.org/10.5054/tq.2010.214047>

- Barkaoui, K. (2010b). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, 27(4), 515–535.
<https://doi.org/10.1177/0265532210368717>
- Barkaoui, K. (2010c). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54–74.
<https://doi.org/10.1080/15434300903464418>
- Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay rating processes and rater performance. *Assessment in Education: Principles, Policy & Practice*, 18(3), 279–293.
<https://doi.org/10.1080/0969594X.2010.526585>
- Beckwith, N. E., & Lehmann, D. R. (1975). The importance of halo effects in multi-attribute attitude models. *Journal of Marketing Research*, 12, 265–275.
<https://doi.org/10.2307/3151224>
- Berlak, H. (1992). The need for a new science of assessment. In H. Berlak, F. M. Newmann, E. Adams, D.A. Archbald, T. Burgess, J. Raven & T. A. Romberg (Eds.), *Toward a new science of educational testing and assessment* (pp. 1–22). Albany, NY: State University of New York Press.
- Berry, R. (2008). *Assessment for learning* (Vol. 1). Hong Kong, China: Hong Kong University Press.
- Bintz, W. P. (1991). Staying connected: Exploring new functions for assessment. *Contemporary Education*, 62(4), 307–312. Retrieved from <https://search.proquest.com/docview/1291764284?accountid=12763>
- Bintz, W. P., & Harste, J. C. (1994). Where are we going with alternative assessment? And is it really worth our time? *Contemporary Education*, 66(1), 7–12.

Retrieved from

<https://search.proquest.com/docview/1291708596?accountid=12763>

Black, H. (1986). Assessment for learning. *Assessing Educational Achievement*, 5(1), 7–18.

Black, P. (2004). *Working inside the black box: Assessment for learning in the classroom*. London, UK: Granada Learning.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London, UK: Granada Learning.

Bond, L. A. (1996). Norm-and criterion-ceferenced cesting. ERIC/AE Digest. *Practical Assessment, Research & Evaluation*, 5(2), 1–3. Retrieved from <https://files.eric.ed.gov/fulltext/ED410316.pdf>

Borg, S. (2012). *Current approaches to language teacher cognition research: A methodological analysis*. Bristol, UK: Multilingual Matters.

Borko, H., & Shavelson, R. J. (1990). Teacher decision making. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 311–346). Hillsdale, NJ: Erlbaum.

Bowers, V. A., & Snyder, H. L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. *Proceedings of the Human Factors Society Annual Meeting*, 34(17), 1270–1274. <https://doi.org/10.1177/154193129003401720>

Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice*, 10(1), 35–36. <https://doi.org/10.1111/j.1745-3992.1991.tb00182.x>

- Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, 22(4), 5–12.
<https://doi.org/10.1111/j.1745-3992.2003.tb00139.x>
- Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria, VA: ASCD.
- Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, 67(3), 343–368.
<https://doi.org/10.1080/00131911.2014.929565>
- Brooks, J. G. (1999). *In search of understanding: The case for constructivist classrooms*. Alexandria, VA: ASCD.
- Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12(1), 1–15.
<https://doi.org/10.1177/026553229501200101>
- Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic purposes speaking tasks* (Monograph Series MS 29). Princeton, NJ: ETS.
- Brown, A., & McNamara, T. (2004). 'The devil is in the detail': Researching gender issues in language assessment. *TESOL Quarterly*, 38(3), 524–538.
<https://doi.org/10.2307/3588353>
- Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, 25(4), 587–603. <https://doi.org/10.2307/3587078>
- Bruce, V., Green, P., & Georgeson, M. (1996). *Visual perception: Physiology, psychology and ecology* (3rd ed.). East Sussex, UK: Psychology Press.
- Bruner, J. S. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.

- Bryman, A. (2006). Integrating quantitative and qualitative research: how is it done? *Qualitative Research*, 6(1), 97–113. <https://doi.org/10.1177/1468794106058877>
- Bryman, A. (2008). Why do researchers integrate/combine/mesh/blend/mix/merge/fuse quantitative and qualitative research. In M. M. Bergman (Ed.), *Advances in Mixed Methods Research* (pp. 87–100). London, UK: SAGE.
- Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays*. Paper presented at the NCME Symposium on Automated Scoring, Montreal.
- Butler, Y. G. (2009). How do teachers observe and evaluate elementary school students' foreign language performance? A case study from South Korea. *TESOL Quarterly*, 43(3), 417-444. <https://doi.org/10.1002/j.1545-7249.2009.tb00243.x>
- Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, 21(2), 1–43. Retrieved from <http://hdl.handle.net/10125/40655>
- Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28(2), 201–219. <https://doi.org/10.1177/0265532210393704>
- Cargile, A. C., & Giles, H. (1998). Language attitudes toward varieties of English: An American-Japanese context. *Journal of Applied Communication Research*, 26(3), 338–356. <https://doi.org/10.1080/00909889809365511>
- Carless, D. (2015). Exploring learning-oriented assessment processes. *Higher Education*, 69(6), 963–976. <https://doi.org/10.1007/s10734-014-9816-z>
- Carless, D. (Ed.) (2010). *Classroom assessment in policy context (Hong Kong)*. Retrieved from http://web.edu.hku.hk/f/acadstaff/412/2010_Classroom-assessment-in-the-Hong-Kong-policy-context.pdf

- Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics*, 11(2), 207–241. Retrieved from <https://escholarship.org/uc/item/4dw4z8rt>
- Carroll, B. (1991). Response to Don Porter's paper: 'Affective factors in language testing'. In J. C. Anderson, B. North, & B. Council (Eds.) *Language Testing in the 1990s: The Communicative Legacy*, (p. 41–45). London, UK: MacMillan.
- Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgement and English language speaking proficiency. *World Englishes*, 24(3), 383–391. <https://doi.org/10.1111/j.0083-2919.2005.00419.x>
- Cherry, K. (2017). *Gestalt laws of perception organisation*. Retrieved <https://www.verywellmind.com/gestalt-laws-of-perceptual-organization-2795835>
- Chung, Y. B. (2002). Career decision-making self-efficacy and career commitment: Gender and ethnic differences among college students. *Journal of Career Development*, 28(4), 277–284. <https://doi.org/10.1023/A:1015146122546>
- Cizek, G. J., Fitzgerald, S. M., & Rachor, R. A. (1995). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment*, 3(2), 159–179. https://doi.org/10.1207/s15326977ea0302_3
- Clark, C. M., & Peterson, P. L. (1984). Teachers' thought processes. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 255 - 296). New York, NY: Macmillan.
- Clough, M. P., Berg, C. A., & Olson, J. K. (2009). Promoting effective science teacher education and science teaching: A framework for teacher decision-making. *International Journal of Science and Mathematics Education*, 7(4), 821–847. <https://doi.org/10.1007/s10763-008-9146-7>

- Colton, A. B., & Sparks-Langer, G. M. (1993). A conceptual framework to guide the development of teacher reflection and decision making. *Journal of Teacher Education, 44*(1), 45–54. <https://doi.org/10.1177/0022487193044001007>
- Connolly, S., Klenowski, V., & Wyatt-Smith, C. M. (2012). Moderation and consistency of teacher judgement: Teachers' views. *British Educational Research Journal, 38*(4), 593–614. <https://doi.org/10.1080/01411926.2011.569006>
- Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation, 13*(5), 401–434. <https://doi.org/10.1080/13803610701728311>
- Croninger, R. G., Rice, J. K., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review, 26*(3), 312–324. <https://doi.org/10.1016/j.econedurev.2005.05.008>
- Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31–51. <https://doi.org/10.1177/026553229000700104>
- Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing, 8*(2), 73–83. [https://doi.org/10.1016/S1075-2935\(02\)00047-8](https://doi.org/10.1016/S1075-2935(02)00047-8)
- Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67–96. <https://doi.org/10.1111/1540-4781.00137>
- Cumming, J., & Maxwell, G. S. (2004). Assessment in Australian schools: Current practice and trends. *Assessment in Education, 11*(1), 89–108. Retrieved from

https://cmap.helsinki.fi/rid=1G5ND2PSC-F1RN00-1VT/assessment_Australianschools.pdf

- Cummins, J. (1996). *Negotiating identities: Education for empowerment in a diverse society*. Ontario, CA: California Association for Bilingual Education.
- Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing, 21*(3), 305–334. <https://doi.org/10.1191/0265532204lt286oa>
- Davison, C. (2007). Views from the chalkface: English language school-based assessment in Hong Kong. *Language Assessment Quarterly, 4*(1), 37–68. <https://doi.org/10.1080/15434300701348359>
- Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly, 43*(3), 393–415. <https://doi.org/10.1002/j.1545-7249.2009.tb00242.x>
- Davison, C., & Michell, M. (2014). EAL assessment: What do Australian teachers want? *TESOL in Context, 24*(2), 51–72. Retrieved from <https://search.informit.com.au/fullText;dn=949508765515560;res=IELHSS>
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford, UK: Oxford University Press.
- Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research: Construction, administration, and processing*. New York, NY: Routledge.
- Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly, 38*(1), 78–103. <https://doi.org/10.1598/RRQ.38.1.4>
- Drucker, M. J. (2003). What reading teachers should know about ESL learners. *The Reading Teacher, 57*(1), 22–29. <https://doi.org/10.1598/RRQ.38.1.4>

- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, 2(3), 197–221.
https://doi.org/10.1207/s15434311laq0203_2
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–185.
<https://doi.org/10.1177/0265532207086780>
- Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, 9(3), 270–292.
<https://doi.org/10.1080/15434303.2011.649381>
- Esterberg, K.G. (2002) *Qualitative Methods in Social Research*. New York, NY: McGraw–Hill.
- Edwards, J. R. (1982). Language attitudes and their implications among English speakers. In E. B. Ryan & H. Giles (Eds.), *Attitudes toward language variation: Social and applied contexts* (pp. 20–33). London, UK: Edward Arnold.
- Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, 37(3), 313–326.
<https://doi.org/10.1111/j.1467-1770.1987.tb00573.x>
- Fonow, M. M., & Cook, J. A. (Eds.). (1991). *Beyond methodology: Feminist scholarship as lived research*. Bloomington, IN: Indiana University Press

- Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, 19(4), 25–42.
<https://doi.org/10.1257/089533005775196732>
- Freire, P. (2000). *Pedagogy of the oppressed*. New York, NY: Bloomsbury Academic.
- Fry, G., Chantavanich, S., & Chantavanich, A. (1981). Merging quantitative and qualitative research techniques: Toward a new research paradigm. *Anthropology & Education Quarterly*, 12(2), 145–158.
<https://doi.org/10.1525/aeq.1981.12.2.05x1889q>
- Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing*, 20(3), 321–344. <https://doi.org/10.1191/0265532203lt259oa>
- Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *The Modern Language Journal*, 64(4), 428–433.
<https://doi.org/10.1111/j.1540-4781.1980.tb05218.x>
- Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System*, 28(1), 31–53. [https://doi.org/10.1016/S0346-251X\(99\)00059-7](https://doi.org/10.1016/S0346-251X(99)00059-7)
- Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34(1), 65–87.
<https://doi.org/10.1111/j.1467-1770.1984.tb00996.x>
- Gears, C. (2005). Classroom assessment: Minute by minute, day by day. *Assessment to Promote Learning*, 63(3), 19–24. Retrieved from
<http://facets.edc.org/sites/facets.edc.org/files/classassessmentdaybyday.pdf>
- Gill, M. M. (1994). Accent and stereotypes: Their effect on perceptions of teachers and lecture comprehension. *Journal of Applied Communication Research*, 22(4), 348–361. <https://doi.org/10.1080/00909889409365409>

- Gipps, C. (2012). *Beyond testing: Towards a theory of educational assessment*. London, UK: Falmer Press.
- Gleeson, M., & Davison, C. (2016). A conflict between experience and professional learning: Subject teachers' beliefs about teaching English language learners. *RELC Journal*, 47(1), 43–57. <https://doi.org/10.1177/0033688216631221>
- Good, T. L., & Lavigne, A. L. (2017). *Looking in classrooms*. New York, NY: Routledge.
- Greene, M. (1988). *The dialectic of freedom*. New York, NY: Teachers College Press
- Gu, Y. (2014). The unbearable lightness of the curriculum: what drives the assessment practices of a teacher of English as a Foreign Language in a Chinese secondary school? *Assessment in Education: Principles, Policy & Practice*, 21(3), 286–305. <https://doi.org/10.1080/0969594X.2013.836076>
- Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning*, 41(1), 1–20. <https://doi.org/10.1111/j.1467-1770.1991.tb00674.x>
- Hallström, L. (1993). What's eating Gilbert Grape [Motion Picture]. *United States: Paramount Pictures*.
- Hamp-Lyons, L. (ed.) (1991). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation.
- Hamp-Lyons, L. (2007). The impact of testing practices on teaching. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 487–504). Norwell, MA: Springer.
- Hamp-Lyons, L. (2009). Principles for large-scale classroom-based teacher assessment of English learners' language: An initial framework from school-based

assessment in Hong Kong. *TESOL Quarterly*, 43(3), 524–530.

<https://doi.org/10.1002/j.1545-7249.2009.tb00249.x>

Han, H., Hsu, L.-T. J., & Lee, J.-S. (2009). Empirical investigation of the roles of attitudes toward green behaviors, overall image, gender, and age in hotel customers' eco-friendly decision-making process. *International Journal of Hospitality Management*, 28(4), 519–528.

<https://doi.org/10.1016/j.ijhm.2009.02.004>

Harris, C. R., Jenkins, M., & Glaser, D. (2006). Gender differences in risk assessment: why do women take fewer risks than men? *Judgment and Decision Making*, 1(1), 48. Retrieved from

<https://search.proquest.com/docview/1010959354?accountid=12763>

Harvey, O. J., Hunt, D. E., & Schroder, H. M. (1961). *Conceptual systems and personality organization*. Oxford, UK: Wiley.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>

He, T.-H., Gou, W. J., Chien, Y.-C., Chen, I.-S. J., & Chang, S.-M. (2013). Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment. *Psychological Reports*, 112(2), 469–485.

<https://doi.org/10.2466/03.11.PR0.112.2.469-485>

Heygt, R. V. D., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, 224, 1260–1262.

<https://doi.org/10.1126/science.6539501>

Hipkins, R. (2010a). Learning through moderation: Minding our language. *set: Research Information for Teachers*, 1, 18. Retrieved from

<https://search.informit.com.au/fullText;dn=320463768093668;res=IELNZC>

- Hipkins, R. (2010b). Reflections on being 'labelled' by National Standards. *set: Research Information for Teachers*, 3, 27. Retrieved from <https://search.informit.com.au/fullText;dn=319681183300824;res=IELNZC>
- Hipkins, R., & Robertson, S. (2011). *Moderation and teacher learning: What can research tell us about their interrelationships?*. Retrieved from <http://www.nzcer.org.nz/system/files/moderation-teacher-learning.pdf>
- Hogan, E. A. (1987). Effects of prior expectations on performance ratings: A longitudinal study. *Academy of Management Journal*, 30(2), 354–368. <https://doi.org/10.2307/256279>
- Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System*, 41(3), 770–785. <https://doi.org/10.1016/j.system.2013.07.009>
- Hunter, A., & Brewer, J. (2003). Multimethod research in sociology. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 577-594). Thousand Oaks, CA: Sage.
- Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, 60(2), 237–263. <https://doi.org/10.3102/00346543060002237>
- Janopoulos, M. (1993). Comprehension, communicative competence, and construct validity: holistic scoring from an ESL perspective. In MW Williamson, BA Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 303-322). Cresskill, NJ: Hampton Press.
- Johnson, B., & Christensen, L. (2000). *Educational research: Quantitative and qualitative approaches*. Boston, MA: Allyn & Bacon.

- Johnson, B., & Christensen, L. (2010). *Educational research: Quantitative, qualitative, and mixed approaches*. Thousand Oaks, CA: Sage Publications.
- Johnson, B., & Onwuegbuzie, A. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14–26.
<https://doi.org/10.3102/0013189X033007014>
- Johnson, B., Onwuegbuzie, A., & Turner, L. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112–133.
<https://doi.org/10.1177/1558689806298224>
- Johnson, J. E., & Powell, P. L. (1994). Decision making, risk and gender: Are managers different? *British Journal of Management*, 5(2), 123–138.
<https://doi.org/10.1111/j.1467-8551.1994.tb00073.x>
- Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, 26(4), 485–505.
<https://doi.org/10.1177/0265532209340186>
- Johnston, P. (1989). Constructive evaluation and the improvement of teaching and learning. *The Teachers College Record*, 90(4), 509–528.
- Johnston, P. (1992). *Constructive evaluation of literate activity*. New York, NY: Longman.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144.
<https://doi.org/10.1016/j.edurev.2007.05.002>
- Kachchaf, R., & Solano-Flores, G. (2012). Rater language background as a source of measurement error in the testing of English language learners. *Applied Measurement in Education*, 25(2), 162–177.
<https://doi.org/10.1080/08957347.2012.660366>

- Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review, 30*(7), 865–878.
<https://doi.org/10.1016/j.cpr.2010.03.001>
- Kennedy, M. (1989). Reflection and the problem of professional standards. *Colloquy, 2*(2), 1–6.
- Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187–217. <https://doi.org/10.1177/0265532208101010>
- Klein, G. (1997). The recognition-primed decision (RPD) model: Looking back, looking forward. In C. E. Zsombok & G. Klein (Eds.), *Naturalistic decision making* (pp. 285–292). Mahwah, NJ: Erlbaum.
- Klenowski, V., & Adie, L. E. (2009). Moderation as judgement practice: Reconciling system level accountability and local level practice. *Curriculum Perspectives, 29*(1), 10–28. Retrieved from
https://eprints.qut.edu.au/26164/1/Klenowski_Moderation_as_judgement.pdf
- Klenowski, V., & Wyatt-Smith, C. (2010). Standards-driven reform Years 1–10: Moderation an optional extra? *The Australian Educational Researcher, 37*(2), 21–39. <https://doi.org/10.1007/BF03216920>
- Knight, P. (2001). *A briefing on key concepts: Formative and summative, criterion and norm-referenced assessment*. Heslington, York: LTSN Generic Centre Assessment Series No.7.
- Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81–96. <https://doi.org/10.1016/j.asw.2011.02.003>

- Knoch, U., & Elder, C. (2010). Validity and fairness implications of varying time conditions on a diagnostic test of academic English writing proficiency. *System*, 38(1), 63–74. <https://doi.org/10.1016/j.system.2009.12.006>
- Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26–43. <https://doi.org/10.1016/j.asw.2007.04.001>
- Koffka, K. (1922). Perception: and introduction to the Gestalt-theorie. *Psychological Bulletin*, 19, 531–585. <http://dx.doi.org/10.1037/h0072422>
- Koffka, K. (1935). *Principles of Gestalt psychology*. London, UK: Lund Humphries.
- Koffka, K. (2013). *Principles of Gestalt psychology* (Vol. 44). New York, NY: Routledge.
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3–31. <https://doi.org/10.1191/0265532202lt218oa>
- Krathwohl, D. (2004). *Methods of educational and social science research*. Long Grove, IL: Waveland Press, Inc.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests. A teacher's book*. New York, NY: McGraw-Hill.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3), 259–284. <https://doi.org/10.1080/01638539809545028>
- Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399–418. <https://doi.org/10.1111/j.1745-3984.2011.00152.x>

- Leung, C. (2005). Classroom teacher assessment of second language development: Construct as practice. In E. Hinkel (Ed.), *Handbook of research in second language* (pp.869 - 888). Mahwah, NJ: Lawrence Erlbaum
- Lumley, T. (2002a). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246–276.
<https://doi.org/10.1191/0265532202lt230oa>
- Lumley, T. (2002b). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, 21(16), 2313–2324. <https://doi.org/10.1002/sim.1201>
- Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54–71.
<https://doi.org/10.1177/026553229501200104>
- Lumley, T., & O’Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, 22(4), 415–437. <https://doi.org/10.1191/0265532205lt303oa>
- Luoma, S. (2004). *Assessing speaking*. Stuttgart, Germany: Ernst Klett Sprachen.
- Lynch, B. K., & McNamara, T. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158–180.
<https://doi.org/10.1177/026553229801500202>
- Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, 36(2), 173–190. <https://doi.org/10.2307/3588329>
- Marshall, C., & Rossman, G. B. (2010). *Designing qualitative research*. Newbury Park, CA: Sage.

- Maxwell, G. S. (2002). *Moderation of teacher judgments in student assessment*. Retrieved from Queensland Curriculum & Assessment Authority: https://www.qcaa.qld.edu.au/downloads/publications/research_qscs_assess_report_2.pdf
- Maxwell, G. S. (2006, May). *Quality management of school-based assessments: Moderation of teacher judgments*. Paper presented at the 32nd International Association for Educational Assessment Conference, Singapore.
- Maxwell, G. S. (2010). Moderation of student work by teachers. In B. McGaw, E. Baker, & P. Peterson (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 457–463). Oxford, UK: Elsevier Science.
- McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, 20(1), 20–32. <https://doi.org/10.1111/j.1745-3992.2001.tb00055.x>
- McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, 22(4), 34–43. <https://doi.org/10.1111/j.1745-3992.2003.tb00142.x>
- McMillan, J. H., & Nash, S. (2000, April). *Teacher classroom assessment and grading practices decision making*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- McNamara, T. (1996). *Measuring Second Language Performance*. London, UK: Longman.
- McNamara, T. (2000). *Language Testing*. Oxford, UK: Oxford University Press.

- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333–349.
<https://doi.org/10.1177/026553220101800402>
- Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7*(25), 1–10.
- Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series, 1996*(1).
- Michell, M. (2017, August 22). Personal discussion.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: SAGE.
- Mitchell, S. E. (1996). Institutions, individuals and talk: The construction of identity in fine art. *International Journal of Art & Design Education, 15*(2), 143–154.
<https://doi.org/10.1111/j.1476-8070.1996.tb00661.x>
- Mitchell, V. W., & Walsh, G. (2004). Gender differences in German consumer decision-making styles. *Journal of Consumer Behaviour, 3*(4), 331–346.
<https://doi.org/10.1002/cb.146>
- Morse, J. M. (2003). Principles of mixed methods and multimethod research design. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 189–208). Thousand Oaks, CA: Sage
- Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22*(4), 13–25. <https://doi.org/10.1111/j.1745-3992.2003.tb00140.x>
- National Assessment Program (2017). NAPLAN. Retrieved from
<https://www.nap.edu.au/naplan>

- Newmann, F. M., & Archbald, D. A. (1992). The nature of authentic academic achievement. In H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven & T. A. Romberg (Eds.), *Toward a new science of educational testing and assessment* (pp. 71–83). Albany, NY: State University of New York Press.
- Nitko, A. J. (2001). *Educational Assessment of Students*. Des Moines, IA: Prentice-Hall Order Processing Center.
- NSW Department of Education. (2006). *ESL Scales*. Retrieved from <https://schoolsequella.det.nsw.edu.au/file/3dc2bcf3-9703-4644-ac75-bd40c1baf94e/1/ESL-scales.pdf>
- NSW Department of Education. (2014). *EAL/D Advice*. Retrieved from https://education.nsw.gov.au/policy-library/associated-documents/eald_advice.pdf
- NSW Department of Education. (2017a). *Supporting EAL/D students*. Retrieved from <https://www.det.nsw.edu.au/wellbeing/succeed/supporting-eald-students>
- NSW Department of Education. (2017b). *Planning EAL/D support*. Retrieved from <https://education.nsw.gov.au/teaching-and-learning/curriculum/multicultural-education/english-as-an-additional-language-or-dialect/planning-eald-support>
- O'Loughlin, K. (2000). *The impact of gender in the IELTS oral interview*. Retrieved from IELTS website: https://www.ielts.org/-/media/research-reports/ielts_rr_volume03_report1.ashx
- O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing*, 19(2), 169–192. <https://doi.org/10.1191/0265532202lt226oa>
- Onwuegbuzie, A., & Johnson, B. (2006). The validity issue in mixed research. *Research in the Schools*, 13(1), 48–63.

- Opdenakker, R. (2006, September). Advantages and disadvantages of four interview techniques in qualitative research. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (Vol. 7, No. 4).
- Ouazad, A. (2008). Assessed by a teacher like me: Race, gender, and subjective evaluations. *International Journal of Art & Design Education* (INSEAD Working Paper No. 2008/57/EPS). Retrieved from <https://ssrn.com/abstract=1267109>
- Page, E. B. (1966). The imminence of ... grading essays by computer. *Phi Delta Kappan*, 47(5), 238–243.
- Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). New York, NY: Routledge.
- Parr, J. M., & Timperley, H. S. (2008). Teachers, schools and using evidence: Considerations of preparedness. *Assessment in Education: Principles, Policy & Practice*, 15(1), 57–71. <https://doi.org/10.1080/09695940701876151>
- Piaget, J. (1970). *Science of education and the psychology of the child*. (D. Coltman, Trans.). Oxford, UK: Orion.
- Popham, J. W. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership*, 62(1), 82.
- Popham, J. W. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, 46(4), 265–273. <https://doi.org/10.1080/08878730.2011.605048>
- Popham, J. W. (2014). *Classroom assessment: What teachers need to know*. Boston, CA: Pearson.

- Porter, D., & Hang, S. S. (1991). Sex, status and style in the interview. *The Dolphin, 21*, 117–128.
- Powell, M., & Ansic, D. (1997). Gender differences in risk behaviour in financial decision-making: An experimental analysis. *Journal of Economic Psychology, 18*(6), 605–628. [https://doi.org/10.1016/S0167-4870\(97\)00026-3](https://doi.org/10.1016/S0167-4870(97)00026-3)
- Pryor, J., & Lubisi, C. (2002). Reconceptualising educational assessment in South Africa—testing times for teachers. *International Journal of Educational Development, 22*(6), 673–686. [https://doi.org/10.1016/S0738-0593\(01\)00034-7](https://doi.org/10.1016/S0738-0593(01)00034-7)
- Punch, K. F. (2009). *Introduction to research methods in education*. London, UK: SAGE.
- Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing, 18*(4), 429–462.
- Resnick, L. B., & Klopfer, L. E. (1989). *Toward the thinking curriculum: Current cognitive research. 1989 ASCD Yearbook*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18–39. <https://doi.org/10.1016/j.asw.2010.01.003>
- Roderick, J. A. (1991). *Context-responsive approaches to assessing children's language*. Urbana, IL: National Council of Teachers of English.
- Roxas, M. L., & Stoneback, J. Y. (2004). The importance of gender across cultures in ethical decision-making. *Journal of Business Ethics, 50*(2), 149–165. <https://doi.org/10.1023/B:BUSI.0000022127.51047.ef>

- Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, 5(1), 77–84.
<https://doi.org/10.1080/0969595980050104>
- Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium* (Vol. 9, pp. 129–152). Orlando, FL: Cambridge University Press.
- Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465–493. <https://doi.org/10.1177/0265532208094273>
- Schofield, J. W. (2002). Increasing the generalizability of qualitative research. In M. Huberman & M. B. Miles (Eds.), *The qualitative researcher's companion* (pp. 171–203). Thousand Oaks, CA: SAGE.
- Schulenberg, J. L. (2007). Analysing police decision-making: Assessing the application of a mixed-method/mixed-model research design. *International Journal of Social Research Methodology*, 10(2), 99–119.
<https://doi.org/10.1080/13645570701334050>
- Scottish Government. (2005). *Assessment is for Learning Programme Information Sheet*. Retrieved from
<http://www.scotland.gov.uk/publications/2005/09/20105413/54156>
- Sessions, R. (1995, November). *Education is a gift, not a commodity*. Paper presented at the National Conference of the Community Colleges Humanities Association, Washington, DC.
- Shavelson, R. J. (1973). The basic teaching skill: decision making. *Journal of Teacher Education*, 24(2), 144–151.

- Shepard, L. A. (1997). *Measuring achievement: What does it mean to test for robust understanding?* Paper presented at the William H. Angoff Memorial Lecture Series, Princeton, NJ.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4–14. <https://doi.org/10.3102/0013189X029007004>
- Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching*. Washington, DC: AERA.
- Shi, L. (2001). Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, 18(3), 303–325. <https://doi.org/10.1177/026553220101800303>
- Shine, K. (2015). Are Australian teachers making the grade? A study of news coverage of NAPLAN testing. *Media International Australia*, 154(1), 25–33. <https://doi.org/10.1177/1329878X1515400105>
- Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, 76(1), 27–33. <https://doi.org/10.1111/j.1540-4781.1992.tb02574.x>
- Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1–23. <https://doi.org/10.17763/haer.57.1.j463w79r56455411>
- Smith, D. B. (1996). Teacher decision making in the adult ESL classroom. In D. Freeman & J. C. Richards (Eds.), *Teacher learning in language teaching* (pp. 197–216). New York, NY: Cambridge University Press.
- Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, 22(4), 26–33. <https://doi.org/10.1111/j.1745-3992.2003.tb00141.x>

- Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, 83(10), 758–765. <https://doi.org/10.1177/003172170208301010>
- Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning: Doing it right—using it well*. Portland, OR: ETS Assessment Training Institute.
- Stoll, L., & Bolam, R. (2005). Developing leadership for learning communities. In M. Coles & G. Southworth (Eds.), *Developing leadership: Creating the schools of tomorrow* (pp. 50–64). Maidenhead, UK: Open University Press.
- Suchy, Y., & Holdnack, J. A. (2013). Assessing Social Cognition Using the ACS for WAIS–IV and WMS–IV. In *WAIS-IV, WMS-IV, and ACS* (pp. 367-406). <https://doi.org/10.1016/B978-0-12-386934-0.00008-0>
- Sweedler-Brown, C. O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing*, 2(1), 3–17. [https://doi.org/10.1016/1060-3743\(93\)90003-L](https://doi.org/10.1016/1060-3743(93)90003-L)
- Tauroza, S., & Luk, J. (1997). Accent and second language listening comprehension. *RELC Journal*, 28(1), 54–71. <https://doi.org/10.1177/003368829702800104>
- Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal*, 60(1), 51–60. <https://doi.org/10.1093/elt/cci081>
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Los Angeles, CA: SAGE.
- Teddlie, C., & Tashakkori, A. (2011). Mixed methods research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (pp. 285-299). Thousand Oaks, CA: SAGE.

- Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, 44(2), 307–336. <https://doi.org/10.1111/j.1467-1770.1994.tb01104.x>
- Thomas, R. M. (2003). *Blending qualitative and quantitative research methods in theses and dissertations*. Thousand Oaks, CA: SAGE.
- Thompson, G. (2014). NAPLAN, MySchool and accountability: Teacher perceptions of the effects of testing. *International Education Journal: Comparative Perspectives*, 12(2), 62-84.
- Timperley, H. (2008). *Teacher professional learning and development* (Educational Practices Series–18). Brussels, Belgium: International Academy of Education & International Bureau of Education.
- Tyndall, B., & Kenyon, D. M. (1996). Validation of a new holistic rating scale using Rasch multifaceted analysis. In A. H. Cumming & R. Berwick (Eds.), *Validation in Language Testing* (pp. 39–57). Clevedon, UK: Multilingual Matters.
- Van Den Haak, M., De Jong, M., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology*, 22(5), 339–351. <https://doi.org/10.1080/0044929031000>
- Van Manen, M. (1977). Linking ways of knowing with ways of being practical. *Curriculum Inquiry*, 6(3), 205–228. <https://doi.org/10.1080/03626784.1977.11075533>
- Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex Publishing Corporation.

- Venkatesh, V., Morris, M. G., & Ackerman, P. L. (2000). A longitudinal field investigation of gender differences in individual technology adoption decision-making processes. *Organizational Behavior and Human Decision Processes*, 83(1), 33–60. <https://doi.org/10.1006/obhd.2000.2896>
- Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, 138(6), 1172–1217. <http://dx.doi.org/10.1037/a0029333>
- Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., & van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin*, 138(6), 1218. <https://doi.org/10.1037/a0029334>
- Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11(2), 197–223. <https://doi.org/10.1177/026553229401100206>
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–287. <https://doi.org/10.1177/026553229801500205>
- Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6(2), 145–178. [https://doi.org/10.1016/S1075-2935\(00\)00010-6](https://doi.org/10.1016/S1075-2935(00)00010-6)
- Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
- Wertheimer, M. (1912). Experimental studies on the seeing of motion. *Psychologia*, 61, 161–265.

- Wertheimer, M. (1923). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 71–88). London, UK: Routledge & Kegan Paul.
- Wertheimer, M. (1938). The general theoretical situation. In W. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 12–16). London, UK: Routledge & Kegan Paul.
- Wertheimer, M. (2012). Experimental studies on seeing motion. In L. Spillmann (Ed.), *On perceived motion and figural organization* (pp. 1–92). Cambridge, MA: The MIT Press.
- Westerman, D. A. (1991). Expert and novice teacher decision making. *Journal of Teacher Education*, 42(4), 292–305.
<https://doi.org/10.1177/002248719104200407>
- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305–319.
<https://doi.org/10.1177/026553229301000306>
- Wilén, W., Bosse, M. I., Hutchison, J., & Kindsvatter, R. (2004). *Dynamics of effective secondary teaching* (5th ed.). Boston, MA: Allyn and Bacon.
- William, D. (2001). An overview of the relationship between assessment and the curriculum. In D. Scott (Ed.), *Curriculum and assessment* (Vol. 1, pp. 165–181). Westport, CT: Ablex Publishing.
- Williamson, M. M., & Huot, B. A. (1993). *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.
- Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, 47(4), 762–789. <https://doi.org/10.1002/tesq.73>

- Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231–252.
<https://doi.org/10.1177/0265532212456968>
- Wolf, D. P. (1992). Assessment as an episode of learning. *Assessment Update, 4*(1), 5–14. <https://doi.org/10.1002/au.3650040105>
- Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *Journal of Effective Teaching, 7*(1), 3–14.
- Wolfe, E. W., Kao, C.-W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4), 465–492.
<https://doi.org/10.1177/0741088398015004002>
- Wyatt-Smith, C. (1999). Reading for assessment: How teachers ascribe meaning and value to student writing. *Assessment in Education: Principles, Policy & Practice, 6*(2), 195–223. <https://doi.org/10.1080/09695949992874>
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170. <https://doi.org/10.1177/0265532209349465>
- Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT [™] speaking section and what kind of training helps?* Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2009.tb02188.x/pdf>
- Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222–1255. <https://doi.org/10.1111/j.1467-9922.2011.00667.x>
- Xu, Y., & Liu, Y. (2009). Teacher assessment knowledge and practice: A narrative inquiry of a Chinese college EFL teacher's experience. *TESOL Quarterly, 43*(3), 492–513. <https://doi.org/10.1002/j.1545-7249.2009.tb00246.x>

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501–527.

<https://doi.org/10.1177/0265532214536171>

Zeichner, K., & Liston, D. (1987). Teaching student teachers to reflect. *Harvard Educational Review, 57*(1), 23–49.

<https://doi.org/10.17763/haer.57.1.j18v7162275t1w3w>

Zhang, Y., & Elder, C. (2010). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?

Language Testing, 28(1), 31–50. <https://doi.org/10.1177/0265532209360671>

Appendices

Appendix A Participant Information Statement and Consent Form

<p style="text-align: center;">School of Education University of New South Wales</p>		
<p>PARTICIPANT INFORMATION STATEMENT</p> <p>Variability in Teacher oral English Assessment Decision-making</p> <p>Prof Chris Davison</p>		
<p>The research study is being carried out by the following researchers:</p>		
<i>Role</i>	<i>Name</i>	<i>Organisation</i>
<i>Chief Investigator</i>	<i>Prof Chris Davison</i>	<i>School of Education, UNSW</i>
<i>Co-Investigator/s</i>	<i>Associate Prof Jihyun Lee</i>	<i>School of Education, UNSW</i>
<i>Student Investigator/s</i>	<i>De Van Phung is conducting this study as the basis for the degree of Doctor of Philosophy in Education at The University of New South Wales. This will take place under the supervision of Prof Chris Davison – Head of School and Associate Prof Jihyun Lee – Senior lecturer.</i>	<i>School of Education</i>
<i>Research Funder</i>	<i>This research is being funded by [list the name/s of funding organisation/s].</i>	

What is the research study about?

You are invited to take part in this online research study. You have been invited because you are currently teaching English as an additional language (EAL) to newly arrived students.

To participate in this research study, you need to meet the following inclusion criteria:

- Being an EAL/D teacher or a specialist working with EAL/D students

The research study is aiming to examine how teachers make assessments of the oral English language of EAL students and explore factors which may influence variability in teachers' assessments.

Do I have to take part in this research study?

This Participant Information Statement tells you about the research study. It explains the research tasks involved. Knowing what is involved will help you decide if you want to take part in the research.

Please read this information carefully. Before deciding whether or not to take part, you might want to talk about it with a relative or friend.

Participation in this research is voluntary. If you don't wish to take part, you don't have to. Your decision will not affect your relationship with The University of New South Wales *and your school*;

What does participation in this research require, and are there any risks involved?

If you decide to take part in the research study, you will be asked to complete an online questionnaire, which will ask you questions about your demographical information such as your gender, language backgrounds, experience and level of education. We expect this activity to take up to 5 minutes.

Deciding to take part in the research study, you will be also asked to carry out an assessment activity, which will ask you to assess a set of students' speaking samples using given assessment rubrics and assessment templates. You will be sent a link to assessment package including a set of videos clips, the assessment rubrics and assessment templates via email. The assessment activity is designed to be done at your convenience and is expected to be completed within 14 days. It should take no longer than 45 minutes of your time.

It is one of the purposes of the research in that information related to your assessment will be correlated with other factors; you will be identified by your email. However, throughout the study, I will use pseudonyms to refer to you as individuals and the institutions where the study takes place. I will not include any information that may identify you in any academic discussions and publications arising from the study or in the thesis.

Will I be paid to participate in this project?

There are no costs associated with participating in this research study, nor will you be paid.

What are the possible benefits to participation?

We hope to use information we get from this research study to benefit others who are teaching and assessing English as a second/additional/foreign language in classroom contexts. We will also give you feedback about your assessment as to how consistent it is with other EAL teachers.

What will happen to information about me?

By clicking on the 'I agree' button you consent to the research team collecting and using information from the questionnaire you complete for the research study. We will keep your data for 7 years in a secure location and no identifiable information will be seen by anyone except the researcher.

It is anticipated that the results of this research study will be published and/or presented in a variety of forums. In any publication and/or presentation, information will be provided in such a way that your research findings may be published, but you will not be individually identifiable in these publications.

Any information obtained in connection with this research study that can identify you will remain confidential. This project will use an external site to create, collect and analyse data collected in a questionnaire format. The site we are using is

www.surveymonkey.com if you agree to participate in this study, the responses you provide to the questionnaire will be stored on a host server that is used by the School of Education. No personal information other than your email will be collected in the questionnaire so only your email will be stored as data. Once we have completed our data collection and analysis, we will import the data we collect to the UNSW server. The data on the host server will then be deleted.

You may be invited to take part in the second stage of this study involving a stimulated think-aloud in which you will be asked to verbalise some of your assessment decisions.

How and when will I find out what the results of the research study are?

You have a right to receive feedback about the overall results of this study. You can tell us that you wish to receive feedback by indicating in the last question of the questionnaire. This feedback will be in the form of a lay summary which is two-page long. You will receive this feedback after the study is finished.

What if I want to withdraw from the research study?

Submitting your completed questionnaire is an indication of your consent to participate in the study. You can withdraw your responses if you change your mind about having them included in the study, up to the point that we have analysed and published the results. You can do this by emailing us via de.phung@student.unsw.edu.au.

What should I do if I have further questions about my involvement in the research study?

The person you may need to contact will depend on the nature of your query. If you want any further information concerning this project or if you have any problems which may be related to your involvement in the project, you can contact the following member/s of the research team:

Research Team Contact

Name	De Van Phung
Position	Student investigator
Telephone	0450375559
Email	de.phung@student.unsw.edu.au

If at any stage during the project you become distressed or require additional support from someone not involved in the research please call:

Contact for feelings of distress

Name/Organisation	De Van Phung
Position	Student investigator
Telephone	0450375559

Email	de.phung@student.unsw.edu.au
--------------	--

What if I have a complaint or any concerns about the research study?

If you have any complaints about any aspect of the project, the way it is being conducted, then you may contact:

Complaints Contact

Position	Human Research Ethics Coordinator
Telephone	+ 61 2 9385 6222
Email	humanethics@unsw.edu.au
HC	HC15541
Reference Number	

Consent Form – Participant providing own consent

Declaration by the participant

- I have read the Participant Information Sheet;
- I understand the purposes, study tasks and risks of the research described in the project;
- I have had an opportunity to ask questions and I am satisfied with the answers I have received;
- I freely agree to participate in this research study as described and understand that I am free to withdraw at any time during the project and withdrawal will not affect my relationship with any of the named organisations and/or research team members;
- I understand that I can download a copy of this consent form from <https://research.unsw.edu.au/application-form-templates>.

Print your name and sign

Appendix B Questionnaire

Variability in teacher oral English assessment decision-making

My name is De Van Phung, a PhD student and my supervisors are Prof. Chris Davison, Head of the School of Education and Dr. Ji Hyun Lee, Senior Lecturer in the School of Education, the University of New South Wales. We are conducting a research project “*Variability of Teacher Oral English Assessment Decision-making*”, which aims to explore teachers’ assessments of oral language performance and influential factors. This questionnaire is designed to collect demographic information from you as teachers of English as an Additional Language and Dialect (EALD). Your responses to this questionnaire will be treated with strict confidentiality in a way that you are not individually identified.

Please respond to the questionnaire by choosing an option that best fits your profile. You will be asked to clarify your answer if your answer is “Other” to any of the questions. The questionnaire should be completed in no more than 20 minutes.

Your time spent completing this questionnaire is highly appreciated and essential to the research.

For the purposes of this research project, please provide your email (this can be your work email or personal email) so that your demographic background can be correlated with your assessments.

Should you have any questions regarding the questionnaire or the research project, please do not hesitate to contact me at de.phung@student.unsw.edu.au.

1. What is your age?
 - a. 25 or under
 - b. 26 – 40
 - c. 41 – 55
 - d. 56 or above
2. What is your gender?
 - a. Male
 - b. Female
3. What languages are spoken in your home?
 - a. English
 - b. Other (specify)
4. What language other than language do you use
 - a. in a regular basis? _____
 - b. occasionally? _____
5. What best describes your current teaching position? (Tick all that apply)
 - a. Primary teacher
 - i. EALD specialist
 - ii. General classroom teacher
 - iii. Administrator (specify _____)
 - iv. Other specialist (specify _____)
 - b. Secondary teacher
 - i. EALD specialist
 - ii. Content teacher (specify _____)
 - iii. Other specialist (specify _____)
 - iv. Administrator (specify _____)
6. What kind of school do you teach currently?
 - a. Government

- i. School
 - ii. English language centre
- b. Catholic
- c. Independent
- 7. Do you have a recognised TESOL specialist qualification?
 - a. Yes
 - b. No
 - c. In progress
- 8. How many years have you been teaching English as an additional language or dialect learners?
 - a. 5 or under
 - b. 6 – 10
 - c. 11 – 15
 - d. 16 or above
- 9. What are the main language groups you are currently teaching? Specify _____
- 10. You may be invited to take part in the second stage of the research. Please indicate whether you are willing to be followed up.
 - a. Yes
 - b. No
- 11. What is your email address? This email address should be your only correspondence throughout the research.

Appendix C Performance Task Descriptions Adapted from TEAL

Task Specification_Task 13: Choosing a gift for a character	
Purpose	To assess learner's ability to be involved in an informal interaction and negotiation with peers.
Description	Learners discuss a suitable gift for a character in a novel or film, a gift that will assist or reward the character, at a certain point in the story.
Assumed Knowledge and Description	<p>Content knowledge: <i>Familiarity with a novel or film being studied</i></p> <p>Text type, genre: <i>Collaborative group discussion</i></p> <p>Linguistic structures and features: <i>Making suggestions and giving reasons to support the suggestion.</i> <i>I/we think...</i> <i>We could/we should...</i> <i>How about...</i> <i>...as it...</i> <i>Reporting a choice and justifying reasons for the choice</i> <i>We chose/decided...</i> <i>Because...</i></p> <p>Vocabulary: <i>Relevant to the character, situation and suggestions for gifts arising from these.</i></p>
<p>Purpose and Value of task</p> <p>This task assesses the ability of students to participate in a collaborative discussion with peers, in which they discuss a character and events in a literary work they are familiar with, in order to reach agreement about a suitable gift for a character in the story. It provides assessment information about EAL students' abilities to negotiate with each other and discuss a literary work they have been studying.</p> <p>Contextual Information</p> <p>The Year 10 students in these videos all studied the same film, <i>What's Eating Gilbert Grape?</i> (1993, J & M Entertainment) prior to being given this task. The students had previously discussed the characters in the film, and the idea of giving a gift to a character that would be useful or suitable for them given their personality or circumstances in the film. The students were grouped by similar language level in order to form groups for the video recording of the discussion.</p> <p>In starting the discussion, the students were asked to say a little about the character they had chosen, and what happened to them in the film, before discussing suitable gifts. The students were asked to discuss a number of possible gifts, and to give reasons the gifts might be suitable, or not be suitable, before coming to a decision.</p>	

Task Specification _Task 19-Movie review	
Purpose	To assess student's ability to give a spoken review of a text (print or visual) they have seen or studied, in response to questions about it from a peer or teacher.
Description	Students read a book/view a movie, and then are asked to give a brief spoken report and evaluation of the work, in response to questions from a classmate or teacher.
Assumed Knowledge and Description	<p>Content knowledge: Familiarity with the genre of a book or film review. Familiarity with the chosen book or movie.</p> <p>Text type, genre: Formal review of a literary work (visual or text), including a report on the main features of a literary work (setting, theme, characters), a summary of the plot, relate the plot to themes or issues, a personal response to a literary text, and evaluation of the work.</p> <p>Linguistic structures and features:</p> <p>Use of simple present in describing features of a book or film</p> <p>Use of either present or past tense in describing a summary of plot</p> <p>Use of either present or past tense to describe a reader's/viewer's response to a book or film</p> <p>Use of present modals to make suggestions and recommendations to others about a literary work.</p> <p>Vocabulary: use of adjectives and adverbs to describe settings, characters and events; use of expressions commonly used in discussing novels or films e.g. 'a great read', 'a must-see film', 'a feel-good novel/film'</p>
<p>Purpose and value of the task</p> <p>This task relates to <u>TEAL Writing assessment Task 19 (A book review)</u> and <u>20 (A film review)</u>, and assesses students' capacity to discuss a literary text or movie they have studied. This includes their ability to describe the plot, characters, relevant themes and issues, and provide evaluative comments and a personal response to the work.</p> <p>The language demands of such a review can be complex and varied. A range of present and past tenses can be used in describing the plot, particular events in the work, the characters, themes and issues arising from the text, and in giving a personal response. Some meanings require present tense, particularly the discussion of themes and issues. Recounting the plot and re-telling events in the story, can be achieved by use of either the 'historic present', such as Paikea rides the whale, or past tense, such as Paikea rode</p>	

Task Specification _ Task 19-Movie review

the whale, when the plot is presented as a narrative. While either present or past can be used, there is an expectation of consistency in the use of one main tense, once the retelling has begun, and that the speaker will continue in the same tense. Similarly, characters can be described either in present or past tenses. Present or past tense can be used in giving a personal response to the work, for example, It's alright, or I thought it **was** good. The challenge for EAL learners is to use this range of tenses consistently in acceptable ways in giving a review.

Commentary and context

The students had all recently studied a literary text, which happened to be a film in all three Samples. The student in Sample 1 had studied Edward Scissorhands, (20th Century Fox, 1990), while the students in Samples 2 and 3 had recently studied Whale Rider, (South Pacific Pictures, 2002). The students were asked to present their reviews as a pair activity, rather than a formal presentation in order to give them support in completion of the task. The student in Sample 3 completed his review in a conversation with his teacher, while the other students held conversations with classmates. The students in video Samples 2 and 3 had reference to notes they developed using the Task sheet (see Task implementation) listing key questions in the interviews, which were used by the students asking the questions, more than the student responding in the Samples. Only three samples were obtained for filming. These samples depict a range of responses, from a fairly basic description of the plot and comment on some aspects of the story, through discussions that build on a retell of the plot to relate it to broader themes such as gender roles, or tradition and change, and make evaluative comments about aspects of the work. Despite their still-developing English language skills, the students effectively communicate a range of meanings in their discussions, and demonstrate a range of language functions, including describing events and characters, identifying themes and issues, and evaluating aspects of the films they had viewed.

Task Specification_ Task 21-Job Interview	
Purpose	To assess learner's ability to interact in the context of an interview about themselves and their personal qualities.
Description	Learners role play a job interview for an imaginary job.
Assumed Knowledge and Description	<p>Content knowledge: <i>Familiarity with a job advertisement and criteria, and application for that job.</i></p> <p>Text type, genre: <i>Formal job interview.</i></p> <p>Linguistic structures and features: <i>Ways of describing one's experience capacities and attributes</i> I have (done)... I can... I am able to... I have experience of... Capacities to act in a hypothetical situation.. <i>I would.., I could...</i></p> <p>Vocabulary: <i>Relevant to the type of position involved in the roleplay.</i></p>
<p>Purpose and value of the task.</p> <p>This task involves an interactive and relatively spontaneous performance in which students are interviewed about themselves in relation to a hypothetical job. It assesses several areas of English language use, including the use of simple present tense to talk about themselves, their qualities and attributes (such as 'I am a creative person'), use of the past tense or present perfect to talk relevant experiences (such as I was.. or I have played...etc), modal verbs to talk about the skills they have (such as can or verb phrases such as I am able to..). It also assesses student's abilities to discuss hypothetical events (such as using conditionals (if ... I would.., and ways of expressing modality, such as adverbs like probably, maybe, or modal verbs such as I might or I would...).</p> <p>The situation also requires students to use culturally appropriate ways of talking about themselves in a positive way, without being judged to be over-confident, conceited or to be bragging. Indeed, it is a delicate balance, for interviewees are expected to sound positive about themselves, yet not overly confident of their own abilities. The task also provides teachers with information about their students' fluency and spontaneity in an interview situation, in which they may be 'put on the spot' by unexpected or difficult questions, within predictable parameters.</p> <p>This oral task is related to <u>TEAL writing assessment task 11 Writing a job application</u></p> <p>Context</p>	

Task Specification_Task 21-Job Interview

The five video samples were collected from two groups of students in different schools. In one school (Samples 1, 4, and 5) some Year 9 students were video recorded in the role-play interview with their teacher, towards the end of a unit of work on occupations and applying for jobs. These interviews related to imaginary but ‘real world’ jobs, as the students had been prepared with relevant background knowledge and language. For these students, therefore, this task involved an element of assessment of achievement in learning in the context of the unit of work. In the second school (video samples 2 and 3), some Year 8 students were asked to participate in the role play at short notice, with only a small amount of time between being asked to participate in a role play for a position of drama captain in the school, and a short verbal notification of the topics to be covered in the interview. They were interviewed by a member of the TEAL team, who they had previously met, rather than a class teacher. In this context, the task had a more diagnostic assessment purpose; to identify the students’ current capacities and weaknesses in the oral language relevant to the task.

Commentary

The task elicited varied performances among the students, which illustrate differences in their oral language capacities. However, performance in this type of task is also affected by the students’ personalities, self-confidence, the degree to which they are gregarious or reserved, the nature of their previous experience in the interview situation, and the extent of their knowledge relevant to the job for which they are being interviewed. So, in this context, personal attributes as well as the oral language knowledge and skills of the students affect the performances of the task. Cultural factors can also influence the students’ performances. The task also provides information about the ways in which students may have adapted to the expectation in Australian culture that people can talk about and project a positive (but not overconfident) image of themselves in this sort of situation. For some students, such as in video Sample 1, this does not appear to be a cultural issue, but for some students talking explicitly and positively about themselves may involve moving away from cultural norms in which such behaviour is not seen as appropriate, especially in younger people. The differences in the ways the students in the videos talk about themselves may reflect such cultural factors, as well as idiosyncratic differences between the students.

Appendix D Assessment Criteria

TEAL Oral Assessment Criteria for Task 13: Choosing a Gift for a Character

Level of Performance	Communication	Cultural conventions of language use	Linguistics structures and features				
			Text structure	Grammatical features	Vocabulary	Phonology	Strategies
4	<ul style="list-style-type: none"> Provides a detailed description of attributes of character and identifies suitable gifts Relates reasons for gifts to attributes of the character Conversational partner(s) to clarify ideas and work together to reach agreement Fluent interaction 	<ul style="list-style-type: none"> Supports conversational partners in constructing and participating in the conversation, assists them when they need assistance Uses language to explicitly manage interaction Makes suggestions Expresses, suggestion, agreement, disagreement, and justification for choice Responding to and guiding partner/s 	<ul style="list-style-type: none"> Long turn to describe character or justify choice Spontaneous turn-taking, with some cooperative interruptions Evaluative comment on suggestions 	<ul style="list-style-type: none"> Accurate use of present tense to describe personalities of characters Accurate use of past tense to describe events in the story Appropriate use of modal verbs – <i>we could.., how about if</i> Use of range of logical connectives to give reasons, – <i>so that, because</i> 	<ul style="list-style-type: none"> Wide range of appropriate word choices – depressed, embarrassed mental problem Occasional errors of form – overweighted 	<ul style="list-style-type: none"> Clearly intelligible Clear articulation of phonemes and connection of sounds Very good control over rhythm, stress and intonation 	<ul style="list-style-type: none"> Manages interaction using appropriate interruptions Explicit appeal for partner’s contribution or support – <i>What do you think?</i> Affirmation of partner’s ideas – <i>That’s a good idea!</i> Explicit request for assistance – <i>I don’t know what to do</i> Provision of support by clarification – <i>Do you mean..?</i>

		participation and contributions					
3	<ul style="list-style-type: none"> Provides a detailed description of attributes of the character and identifies suitable gifts Relates reasons for the choice of gift to attributes of the character. Conversational partners work together to reach agreement Fluent interaction 	<ul style="list-style-type: none"> Works collaboratively with partners in turn taking and constructing the conversation Uses language to explicitly structure interaction Expresses suggestion, agreement, disagreement, and justification for choice Responding to partner/s and making contributions Constant eye contact, responding to partner(s) 	<ul style="list-style-type: none"> Longer turns to describe character or justify choice Spontaneous turn taking, in cooperation with partner(s) Suggestions and evaluative responses 	<ul style="list-style-type: none"> Mostly accurate use of present tense to describe personalities of characters Mostly accurate use of past tense Mostly appropriate expression of modality – we could, maybe Use of greater variety of terms in expressing reasons – ...and then..., because if.. 	<ul style="list-style-type: none"> Increased matching of semantic choice and form of word – <i>disability</i> Some errors of word form and expressions – <i>truck is broken child for children furniture.</i> 	<ul style="list-style-type: none"> Intelligible Clear articulation of phonemes and connection of sounds Some errors such as omission of final consonant – <i>book</i> for <i>books</i> Good control over rhythm, stress and intonation 	<ul style="list-style-type: none"> Participates in interaction to reach agreement Explicit appeal for help, request for feedback on own contribution Use of circumlocution – <i>like a chair or something</i> Accepting parts of ideas but rejecting other parts Referring to partner by name
2	<ul style="list-style-type: none"> Provides a description of character and identifies suitable gifts Gives justification for 	<ul style="list-style-type: none"> Turn taking is formalized but not very spontaneous, sometimes signalled only by looking at partner Some use of language relevant to turn taking and 	<ul style="list-style-type: none"> Alternating turns of moderate length Some formulaic 	<ul style="list-style-type: none"> Use of present tense to describe characters Use of past tense to describe events in the story 	<ul style="list-style-type: none"> Clear semantic meaning, but sometimes incorrect 	<ul style="list-style-type: none"> Intelligible, but some noticeable mispronunciation of some sounds – /g/ for /k/ in <i>Becky</i> 	<ul style="list-style-type: none"> Explicitly asks for ideas e.g. <i>What do you think?</i> Taking over from partner when they are stuck

	<ul style="list-style-type: none"> choice of gift Presents own ideas, and responds to partner's ideas Some pauses and hesitation in interaction 	<p>interaction, such as direct use of questions – <i>How about...?</i></p> <ul style="list-style-type: none"> Expresses suggestion, agreement and disagreement Nodding head as back channelling, feedback to conversational partner Constant eye contact with partners, but sometimes looking away from the conversation 	<p>phrases used in signalling shift of turn – <i>What do you think? How about...?</i></p> <ul style="list-style-type: none"> Suggestions with reasons and responses 	<ul style="list-style-type: none"> Some use of modals – we could, we should. Some errors of subject-verb agreement – He take care of him Use of because to give reasons 	<p>forms – <i>mentally sick,</i></p> <ul style="list-style-type: none"> Errors of word choice – <i>stay in his way, at the first.</i> 	<ul style="list-style-type: none"> Impression of separated words, rather than constant flow of speech Problems with some consonant clusters, – /ld/ in old Usually flat intonation, but some variation to show enthusiasm – I think that's a good idea 	<ul style="list-style-type: none"> Some asking of questions to support partners Self-correction of errors – <i>happy...happiness'</i>
1	<ul style="list-style-type: none"> Provides a limited description of the character and identifies suitable gifts Gives short justification for gifts, Exchange of ideas 	<ul style="list-style-type: none"> Turn taking, but often not signalled by language Express suggestion – <i>how about...</i>, agreement – <i>Ok it's a good idea...</i> and disagreement – <i>that is not a good idea.</i> 	<ul style="list-style-type: none"> Alternating turns of moderate length Minimal language used in signalling shift of turn 	<ul style="list-style-type: none"> Sentence and clause construction errors – <i>maybe it a little bit not good idea...</i> Errors in formation of questions – <i>How about you think...</i> 	<ul style="list-style-type: none"> Clear semantic meaning, but sometimes incorrect form of word used – <i>obesity'for</i> 	<ul style="list-style-type: none"> Intelligible pronunciation Noticeable errors in production of some sounds such as /r/ especially in consonant 	<ul style="list-style-type: none"> Looking at partner when unable to continue Use of gesture to assist when struggling for a word Lending support to partner by giving the answer, correcting what

	<ul style="list-style-type: none"> • Frequent pauses and hesitation, searching for ideas or words to use 	<ul style="list-style-type: none"> • Eye contact not maintained, looking in direction of partner more than eye contact, or even looking elsewhere while speaking 	<ul style="list-style-type: none"> • Suggestions with reasons and responses 	<ul style="list-style-type: none"> • Extensive use of present tense, even to retell events of the story • Limited use of modality – <i>maybe</i> • Frequent errors of subject-verb agreement <i>Gilbert takes care...</i> • Use of because to give reasons • Inappropriate use of conjunctions – <i>about</i> 	<ul style="list-style-type: none"> <i>obese, he is loyalty to, for he is loyal to, die for dead</i> • gaps in relevant vocabulary – problem of his mental 	<ul style="list-style-type: none"> clusters-<i>Grape, problem</i> • Omission of final consonants – <i>end of house</i> • Some sounds and words difficult to identify • Relatively flat intonation 	<ul style="list-style-type: none"> partner says, whispering a response, prompting or completing a phrase for partner when partner is ‘stuck’ • Uses circumlocution when word is not known – <i>problem of his mental for intellectual disability</i>
<p>Marked performance level (e.g. 1, 2, 3 or 4): Comments:</p>							

TEAL Oral Assessment Criteria for Task 19: A Book or Movie review: Criteria sheet

Level of performance	Communication	Cultural conventions of language use	Linguistic structures and features				Strategies
			Text structure	Grammatical features	Vocabulary	Phonology	
4	<ul style="list-style-type: none"> Describes the plot in detail, and relates to themes and issues Describes characters and how they illustrate or relate to themes or issues Describes key events and how they relate to themes or issues Relates personal evaluation of the work to elements of the work Relates work to self Comments on elements related to filmography, literary techniques or devices and their impact 	<ul style="list-style-type: none"> Very fluent interaction, responding to questions 	<ul style="list-style-type: none"> Long turns Extended statements Details of text related to themes and issues, and deeper personal responses 	<ul style="list-style-type: none"> Use of a wide range of tenses used appropriate and consistently in expressing different types of meanings Use of additional verb tenses, modals verbs to discuss hypothetical and conditional meanings use of a range of adverbs to express modality and qualify or emphasise <i>probably possibly, actually, etc</i> 	<ul style="list-style-type: none"> Uses and explains a range of specialised terminology from the work Uses terminology related to the themes and issues e.g. <i>gender roles, tradition and change</i> 	<ul style="list-style-type: none"> Clearly intelligible with no problems for audience 	<ul style="list-style-type: none"> Self-sustained presentation with little or no reference to notes or prompts
3	<ul style="list-style-type: none"> Describes the plot and events in details Describes events and their significance, and explains the significance related to themes and issues Gives a personal response, relates elements of the text to self Makes evaluative comment on elements of the work Makes evaluative comment about the work as a whole 	<ul style="list-style-type: none"> Fluent interaction, answering questions 	<ul style="list-style-type: none"> Describing text describing events and relating them to themes Evaluative comments on the aspects of the work 	<ul style="list-style-type: none"> Consistent tense use, either past or present, to describe plot, characters, and make evaluative comments Some use of conditional <i>If I were</i> in relating story to self use of adverbs like <i>done quite well, I'm pretty sure, actually</i> to qualify or emphasize 	<ul style="list-style-type: none"> Uses and explains some specialised terms from the work e.g. <i>taiaha</i> Uses some terminology relevant to issues and themes in the work – <i>gender equality</i> 	<ul style="list-style-type: none"> Clear intelligible pronunciation, though non – standard pronunciation of some words e.g. <i>Maori</i> pronounced as <i>my-ori</i> 	<ul style="list-style-type: none"> Explicit request for assistance, <i>I'm not sure ...</i> Some use of notes
2	<ul style="list-style-type: none"> Describes main characters and significant events of the plot Relates elements of the work to the themes or issues Makes some evaluative comments about elements of the text, such as the believability of a scene using props. Fluent, but some hesitation at times 	<ul style="list-style-type: none"> Interacts, answering questions and providing reasons and explanations Appropriate turn taking, and sharing of ideas 	<ul style="list-style-type: none"> Moderate turns, long turn in describing the plot of the text Discussion of plot, characters themes Simple personal reactions To issues and parts of the text e.g. <i>I liked the story ...</i> 	<ul style="list-style-type: none"> Mostly consistent use of present or past tense to re tell narrative Mixture of present and past tense used in discussing characters Mixture of present and past tenses used to discuss issues and give responses to elements of the text 	<ul style="list-style-type: none"> Uses terminology, names, places, ideas etc Limited range of vocabulary for evaluative comments 	<ul style="list-style-type: none"> Intelligible, sounds clearly articulated 	<ul style="list-style-type: none"> Responds to questions and adds more information Uses gestures to add meaning, including actions depicted in the film May rely on notes

<p>1</p>	<ul style="list-style-type: none"> Identifies main characters and events of the plot Identifies theme or issue in the movie or film Describes a reaction to the text Pauses and hesitations to think about or plan comments 	<ul style="list-style-type: none"> Responds to questions asked by conversational partner Appropriate turn taking and addressing of conversational partner 	<ul style="list-style-type: none"> Questions and answers about plot and main characters Questions and answers about reaction to elements in the work Short to moderate length turns 	<ul style="list-style-type: none"> Inconsistent use of past tense to retell narrative elements of the story Present tense used to describe aspects of characters <i>Because</i> used to give reasons 	<ul style="list-style-type: none"> Uses minimal terminology relevant to the work Limited vocabulary for describing response to the text or aspects of it –<i>I feel nice</i> 	<ul style="list-style-type: none"> Intelligible, but some perceptive errors of production, such as omission of final consonants Some stress or rhythm errors makes speech sound uneven 	<ul style="list-style-type: none"> Requests for clarification – <i>What do you mean?</i> May avoid answering difficult questions May rely extensively on notes
<p>Marked performance level (e.g. 1, 2, 3 or 4): Comments:</p>							

TEAL Oral Assessment Criteria for Task 21: Job interview role play

Level of performance	Communication	Cultural conventions of language use	Linguistic structures and features				Strategies
			Text structure	Grammatical features	Vocabulary	Phonology	
4	<ul style="list-style-type: none"> Participates in role play without hesitation or pauses Describes qualities and skills in general terms more than specific experiences Describes range of potential actions in the job Presents very strong reasons for being employed 	<ul style="list-style-type: none"> Self-assured and confident interaction with interviewer Talks about self with high level of self-assurance 	<ul style="list-style-type: none"> Questions and elaborate answers from student Student describes how they would act in the role 	<ul style="list-style-type: none"> Range of present and past tenses used Use of modals and conditionals and hypothetical situations, e.g. <i>I would recommend I would call..</i> Use of adverbs to add emphasis e.g. <i>I'm really interested</i> 	<ul style="list-style-type: none"> Uses a range of vocabulary for skills and attributes, e.g. <i>really interested in sport, hard-working</i> Uses a range of work place terminology e.g. <i>customer, staff team, part time, work together</i> 	<ul style="list-style-type: none"> Very clear articulation of sounds Appropriate linking of sounds Generally appropriate rhythm, stress and intonation 	<ul style="list-style-type: none"> Avoidance by explicit, but acceptable, declination to answer a question
3	<ul style="list-style-type: none"> Participates in role play with high level of fluency Describes qualities and experience in specific examples, some comments about general attributes Detailed response to hypothetical situation Presents strong reasons for being employed 	<ul style="list-style-type: none"> Generally confident interaction with interviewer Talks about self comfortably 	<ul style="list-style-type: none"> Questions and detailed responses to questions Relates experience to the role 	<ul style="list-style-type: none"> Past and present tenses used Use of modal verbs for hypothetical situations e.g. <i>I would.. I'd first...</i> Adverbs to qualify verbs e.g. <i>I am pretty comfortable</i> 	<ul style="list-style-type: none"> Uses some appropriate terminology for attributes Uses some specific terminology related to the position e.g. <i>role, character, personal trainer</i> 	<ul style="list-style-type: none"> Clear articulation of sounds Mostly appropriate linking of sounds More often than not appropriate rhythm, stress and intonation 	<ul style="list-style-type: none"> Avoidance by limiting or qualifying answer e.g. <i>Yeah, Not sure, No</i> or concluding with <i>I think like that</i>
2	<ul style="list-style-type: none"> Participates in interview with overall fluency Describes qualities and experience mainly in terms of specific examples and strategies Gives response to hypothetical situation Presents plausible reasons for being employed 	<ul style="list-style-type: none"> Generally, interacts with ease in interview, but has difficulty in interaction at some points Able to talk about self-confidently, but hesitation and reservation at points 	<ul style="list-style-type: none"> Questions and some longer responses by student Student provides information to explain answers 	<ul style="list-style-type: none"> Simple present and simple past tenses used appropriately for skills and experience Use of present tense to describe hypothetical events Modality mainly expressed by adverbs, e.g. <i>probably, maybe, rather than verbs.</i> 	<ul style="list-style-type: none"> Uses some terminology for attributes, but not always correct form of the word e.g. <i>open wider</i> for 'open' or 'approachable', <i>On time</i> for <i>arrive on time</i> Uses some work or role related terminology 	<ul style="list-style-type: none"> Most sounds clear, but some omission of sounds e.g. final consonants <i>interes(t) new(s)</i> Some stress and rhythm errors makes speech sound a little stilted, and some syllables difficult to hear 	<ul style="list-style-type: none"> Avoidance by giving a short answer Use of gesture to support meaning

<p>1</p>	<ul style="list-style-type: none"> • Participates in interview, but noticeable points hesitation and loss of fluency • Describes qualities based on specific past experiences • Has difficulty explaining potential or hypothetical actions • Presents reasons for being employed 	<ul style="list-style-type: none"> • Self-conscious, uneasy interaction with interviewer • Talks about self with limited self-confidence or self-consciousness 	<ul style="list-style-type: none"> • Questions and mostly short responses to questions • Few long turns by student 	<ul style="list-style-type: none"> • Use of simple present for skills or attributes • Simple past e.g. <i>I did, I worked</i> for experiences, but not consistent • Use of modal verbs limited to <i>I can</i> 	<ul style="list-style-type: none"> • Uses limited terminology for personal attributes e.g. <i>well-organised honest</i> • Uses limited job related terminology, e.g. <i>Part time</i> 	<ul style="list-style-type: none"> • Intelligible, but some noticeable sounds, stress and rhythm, reflect L1; e.g. <i>singing</i> sounds like 'sin-ging' /sɪŋɪŋ/ • Some sounds or words difficult to recognise 	<ul style="list-style-type: none"> • Avoidance of difficulty by explicit statement (e.g. <i>I don't know in English</i>) or request to teacher to move to next question
<p>Marked performance level (e.g. 1, 2, 3 or 4): Comments:</p>							

Appendix E Guidelines for Think-aloud Protocol and Interview Questions

Stimulus question

Now you have just had a chance to watch the videos again and revisit your work. Could you please explain how you did what you did and why you did that?

Interview Questions

This is a semi-structured interview, so these questions are flexible and can be changed to adapt to the flow of communication. The interviewer is encouraged to modify these questions to ensure correct understanding from teachers. The interviewer is also allowed to ask follow-up questions to obtain sufficient information from the respondents.

1. I am interested in how confident you felt in assessing these students? Can you elaborate?
2. a. When you first listened to the student(s), you didn't have the criteria, so what did you mainly look for in assessing their oral language performance?
b. To what extent did looking at the criteria change your assessment?
c. To what extent did your impression of student strengths influence your assessment?
d. During your assessment did you look for plus points from student strengths OR minus points from his weaknesses? Can you elaborate a bit more?
3. a. To what extent were you aware of any other influences on your assessments? For example, gender bias, language bias?
b. To what extent did previous exposure to students of a specific language group help you to understand students' spoken language, in this case Chinese and Mongolian?
c. All of the students come from different language backgrounds; to what extent do you think this affected your assessment?
d. In the first task, one girl conversed with two boys; one boy with one girl in the second task and one boy with a male teacher in the third task. To what extent do you think gender affected your assessment?
4. a. To what extent do you think your qualifications (or lack of qualifications) helped or hindered you in assessing these students?
b. To what extent did you think your own teaching experiences influenced your assessment? E.g. any difficulties understanding students, interpreting criteria and assessing student performance?
5. To what extent do you think the nature of the task itself influenced your assessment? e.g. which is the easiest, equally difficult....?
6. Do you have any other comments?

Appendix F Participant Information Statement and Consent Form

Teachers	Student	Com	Cul	Text	Gram	Vocab	Phono	Stra
T1	S1	4	4	4	4	4	4	4
T2		3	4	4	4	4	4	4
T3		2	2	3	4	3	3	3
T4		3	2	2	2	2	2	2
T5		2	2	2	2	2	1	2
T6		2	2	2	2	3	3	3
T7		3	3	3	2	2	2	3
T8		3	3	2	2	3	2	3
T9		3	3	3	4	3	4	4
T10		3	2	2	2	2	2	2
T11		2	2	2	2	2	3	2
T12		4	3	2	2	2	2	4
T1	S2	4	4	2	2	3	2	4
T2		3	3	3	3	3	3	3
T3		4	4	3	4	3	3	3
T4		3	4	3	3	3	2	4
T5		4	4	2	3	4	1	4
T6		3	3	2	3	2	2	3
T7		3	3	3	3	3	3	3
T8		2	2	2	3	2	2	1
T9		2	2	3	2	3	2	3
T10		3	4	3	2	3	3	3
T11		2	2	3	3	3	1	2
T12		3	2	2	2	2	1	2
T1	S3	3	4	4	3	3	4	4
T2		3	3	3	3	3	4	3
T3		4	4	4	4	4	4	4
T4		4	3	4	4	3	3	3
T5		4	4	4	3	3	4	4
T6		3	3	3	3	4	4	3
T7		3	4	3	4	3	4	3
T8		3	4	3	4	3	4	4
T9		4	4	4	3	3	4	3
T10		3	3	4	3	3	4	3
T11		2	3	3	3	3	4	4
T12		3	3	3	3	3	3	2